Eric Jonas
Department of Computer Science
University of Chicago
@stochastician
ericj@uchicago.edu

molecular inverse problems

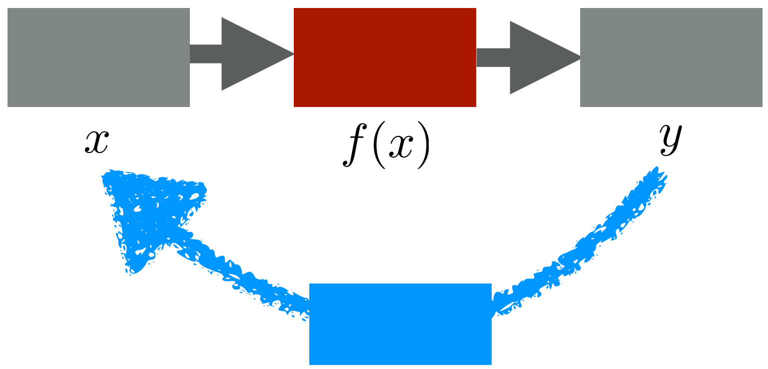
Spectroscopy and RDKit

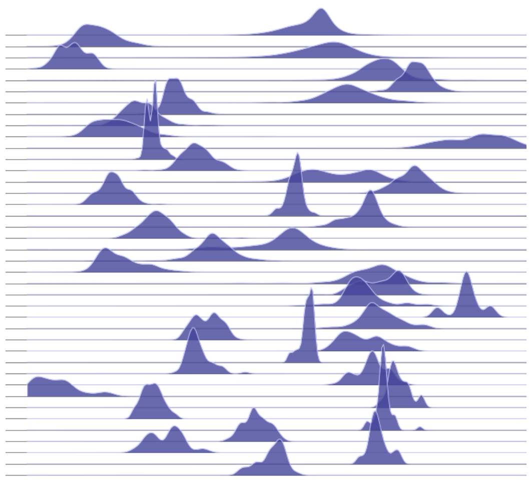
Eric Jonas

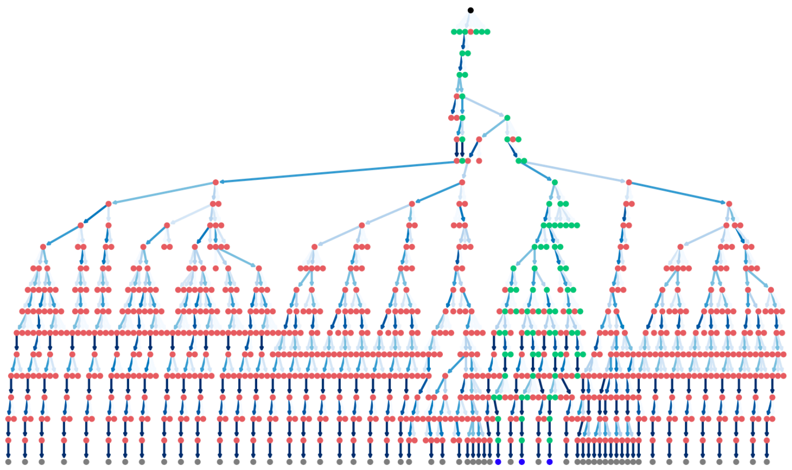
Department of Computer Science
University of Chicago

@stochastician
ericj@uchicago.edu

TODAY





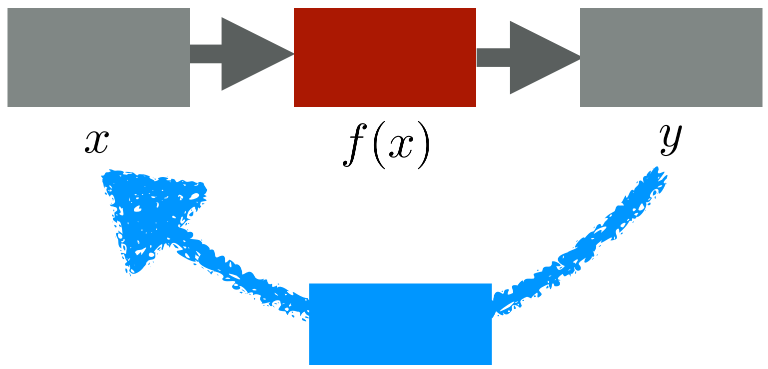


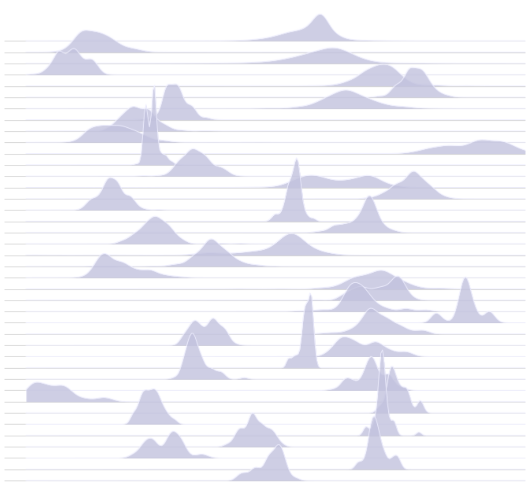
What are
inverse problems?

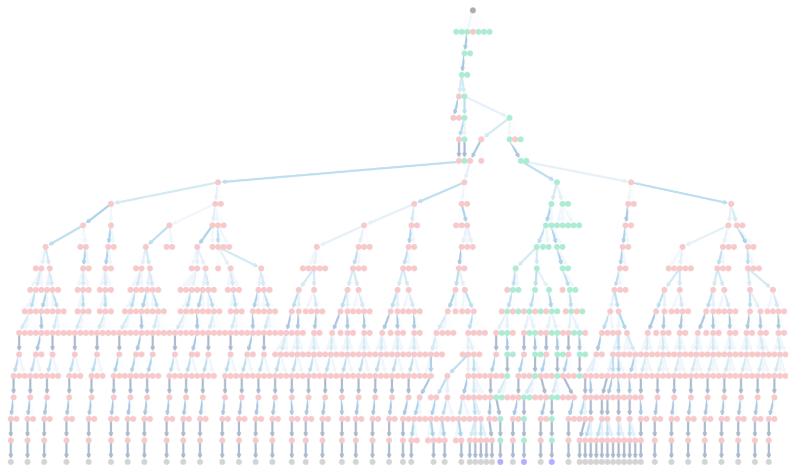
Spectroscopy:
The forward problem

Spectroscopy:
The inverse problem

TODAY







What are
inverse problems?

Spectroscopy:
The forward problem

Spectroscopy:
The inverse problem

What is an

INVERSE PROBLEM?

What is an

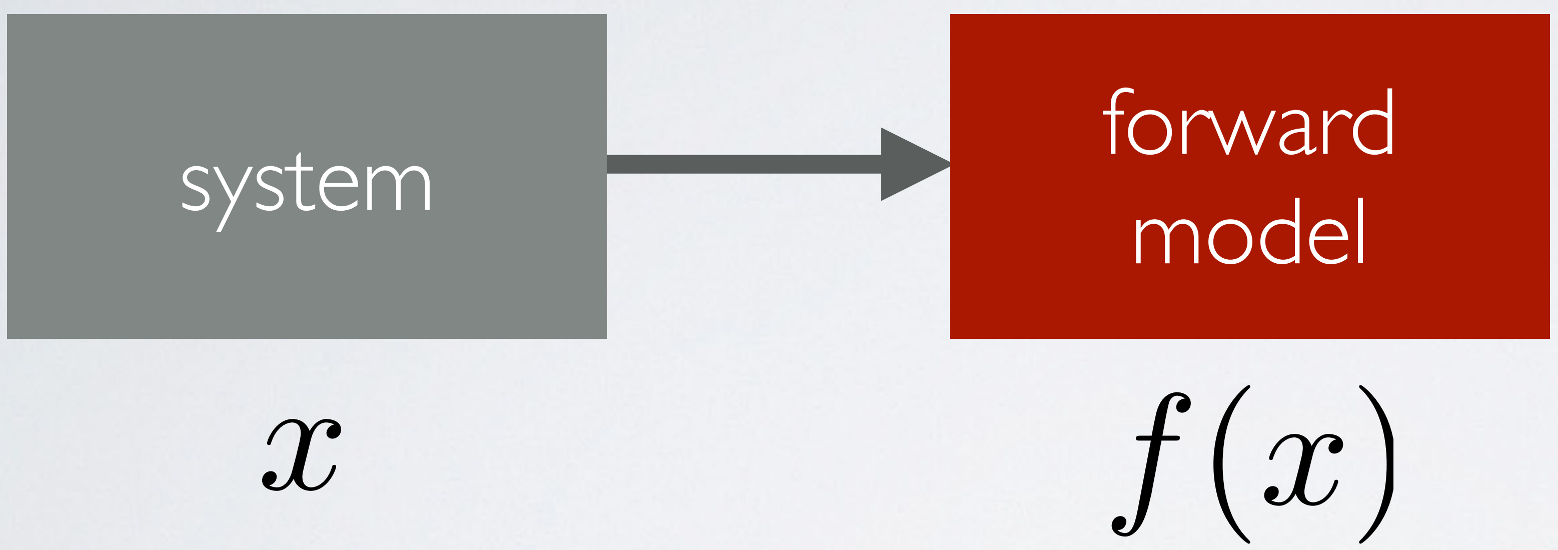
INVERSE PROBLEM?

system

$$x$$

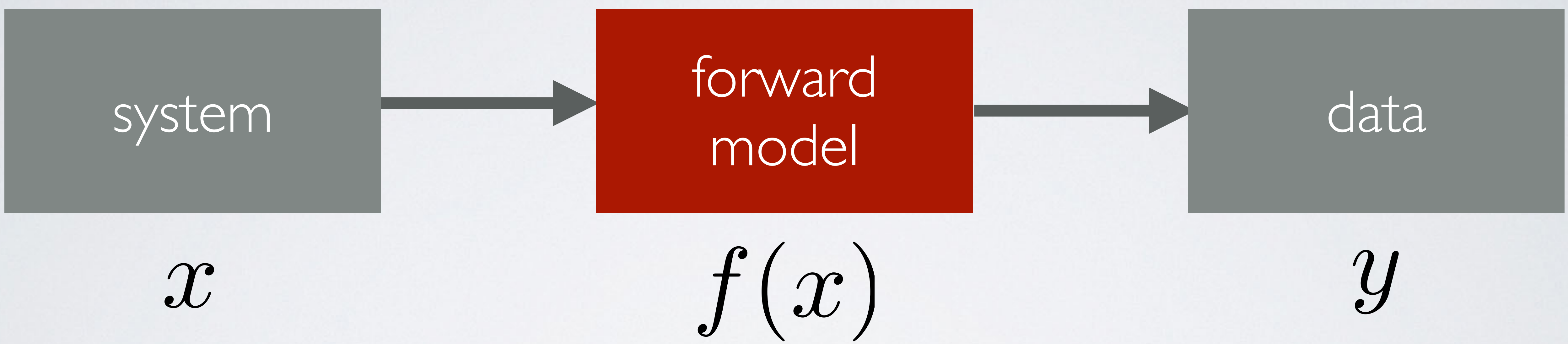
What is an

INVERSE PROBLEM?



What is an

INVERSE PROBLEM?



What is an

INVERSE PROBLEM?

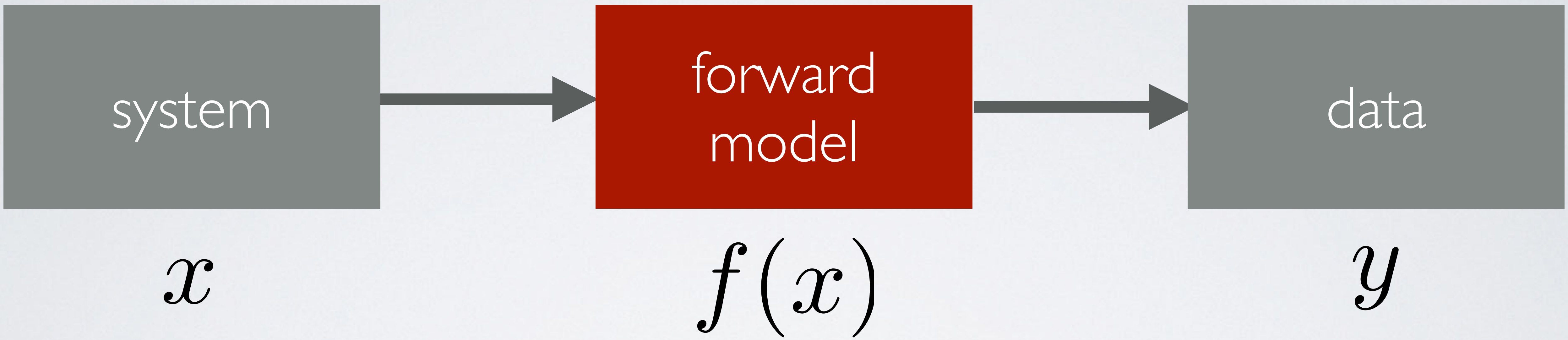
$$x \qquad f(x) \qquad y$$

$$y = f(x) + \epsilon$$

What is an

INVERSE PROBLEM?

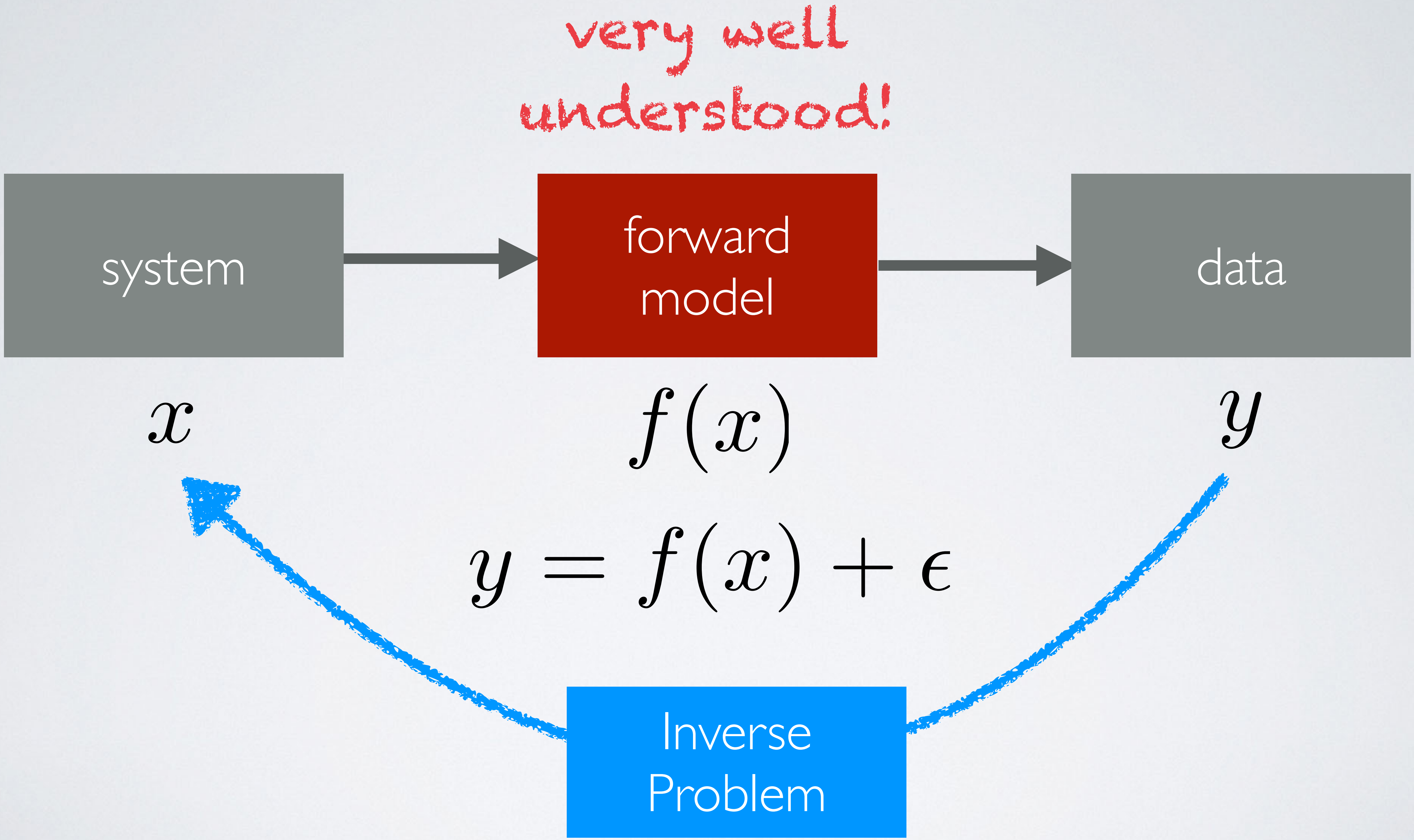
very well
understood!



$$y = f(x) + \epsilon$$

What is an

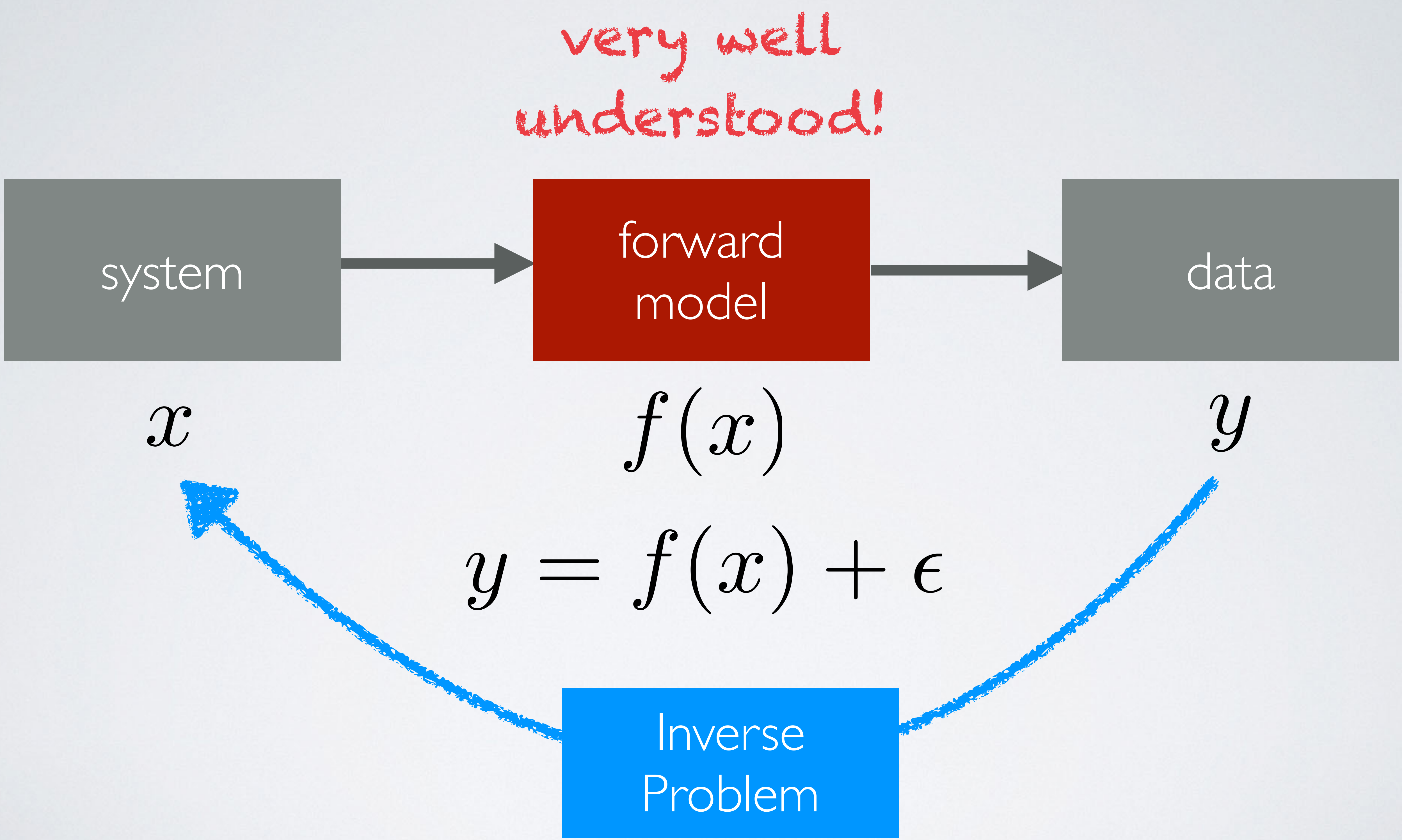
INVERSE PROBLEM?



$$y = f(x) + \epsilon$$

What is an

INVERSE PROBLEM?



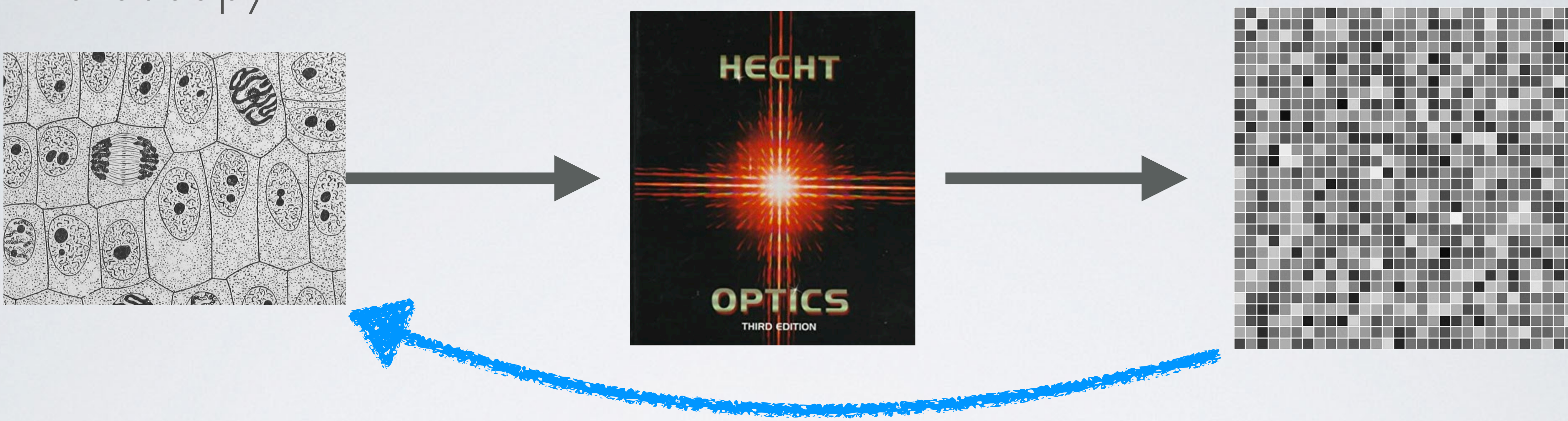
Crucial distinction: **single correct answer**

EXAMPLE PROBLEMS

Microscopy

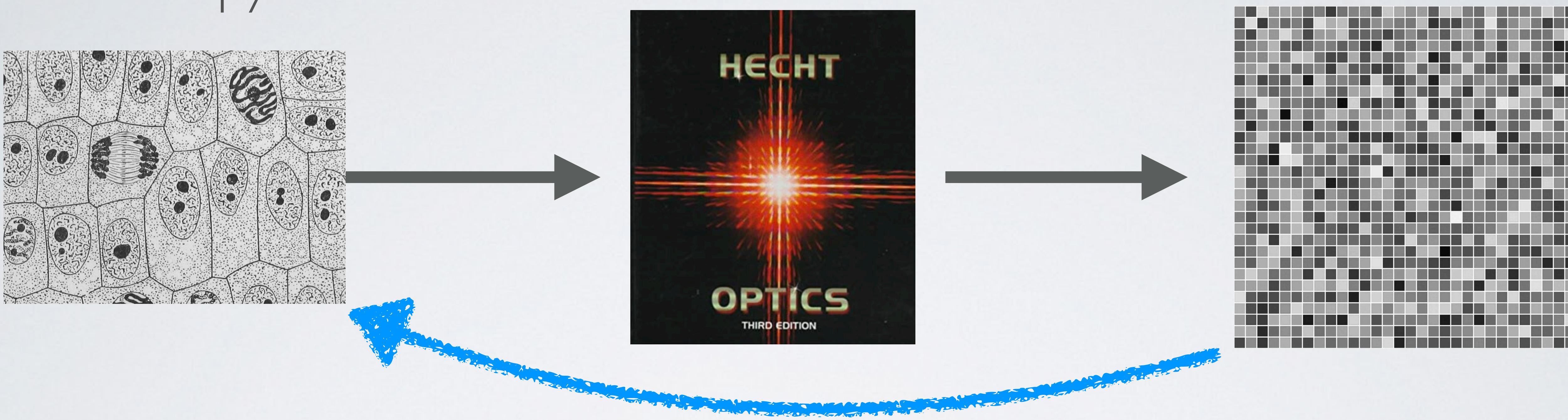
EXAMPLE PROBLEMS

Microscopy



EXAMPLE PROBLEMS

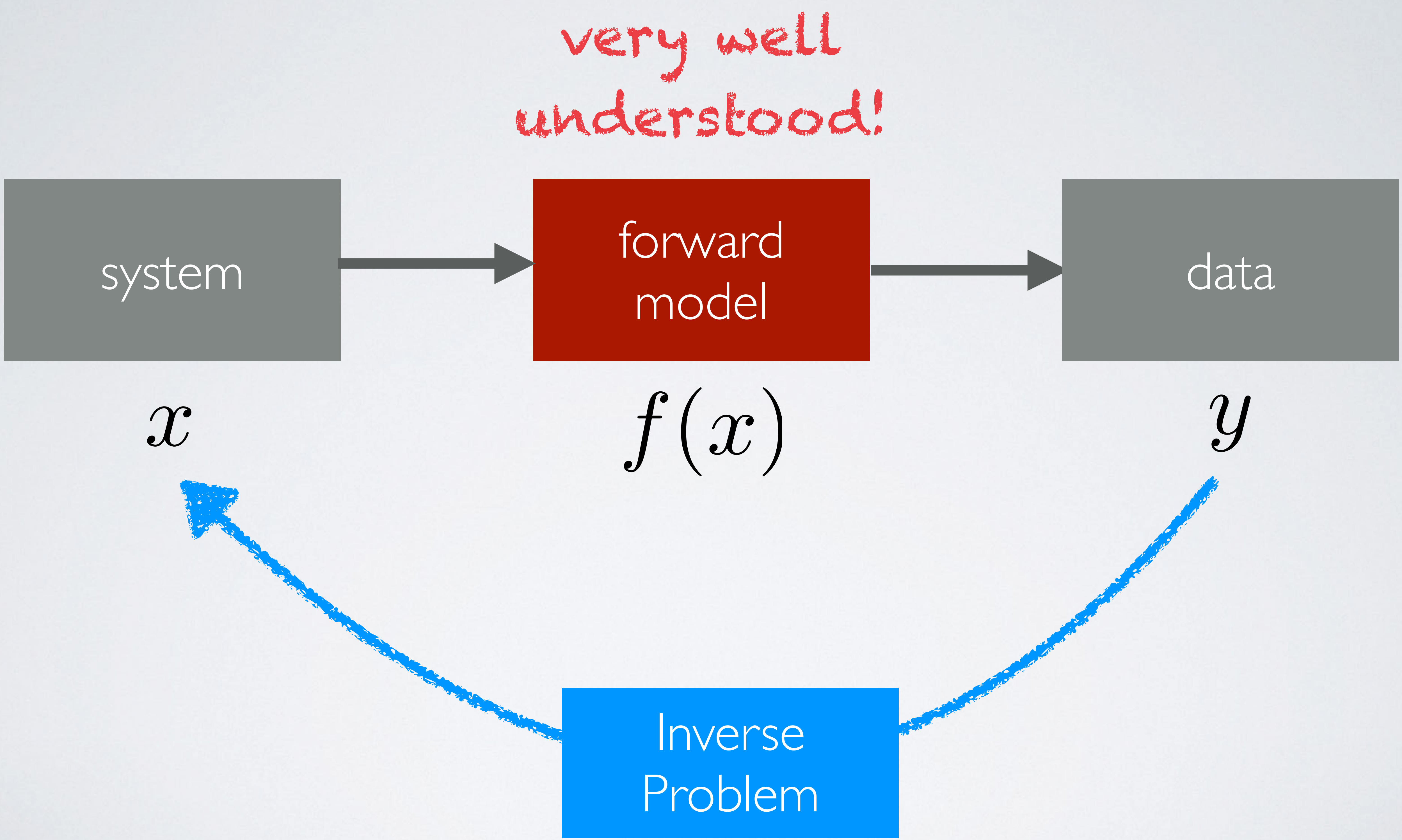
Microscopy



Magnetic Resonance Imaging

Spectroscopy as an

INVERSE PROBLEM



Spectroscopy as an

INVERSE PROBLEM

very well understood!

forward model

data

x

$f(x)$

y

Inverse Problem

Spectroscopy as an

INVERSE PROBLEM

x

data

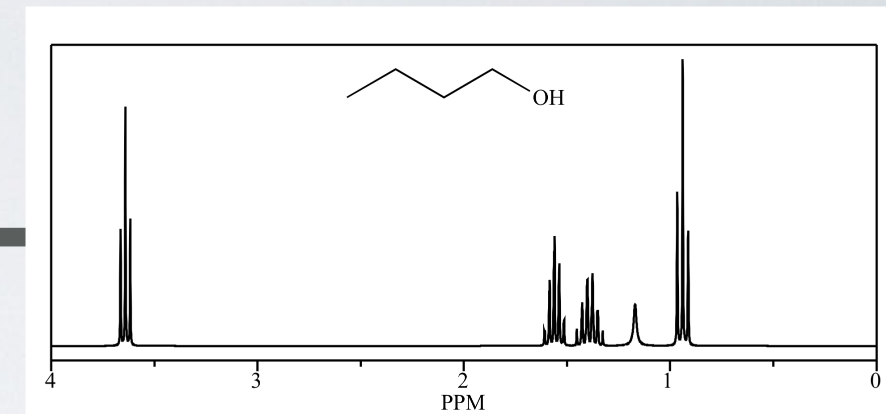
y

Inverse Problem

Spectroscopy as an

INVERSE PROBLEM

x

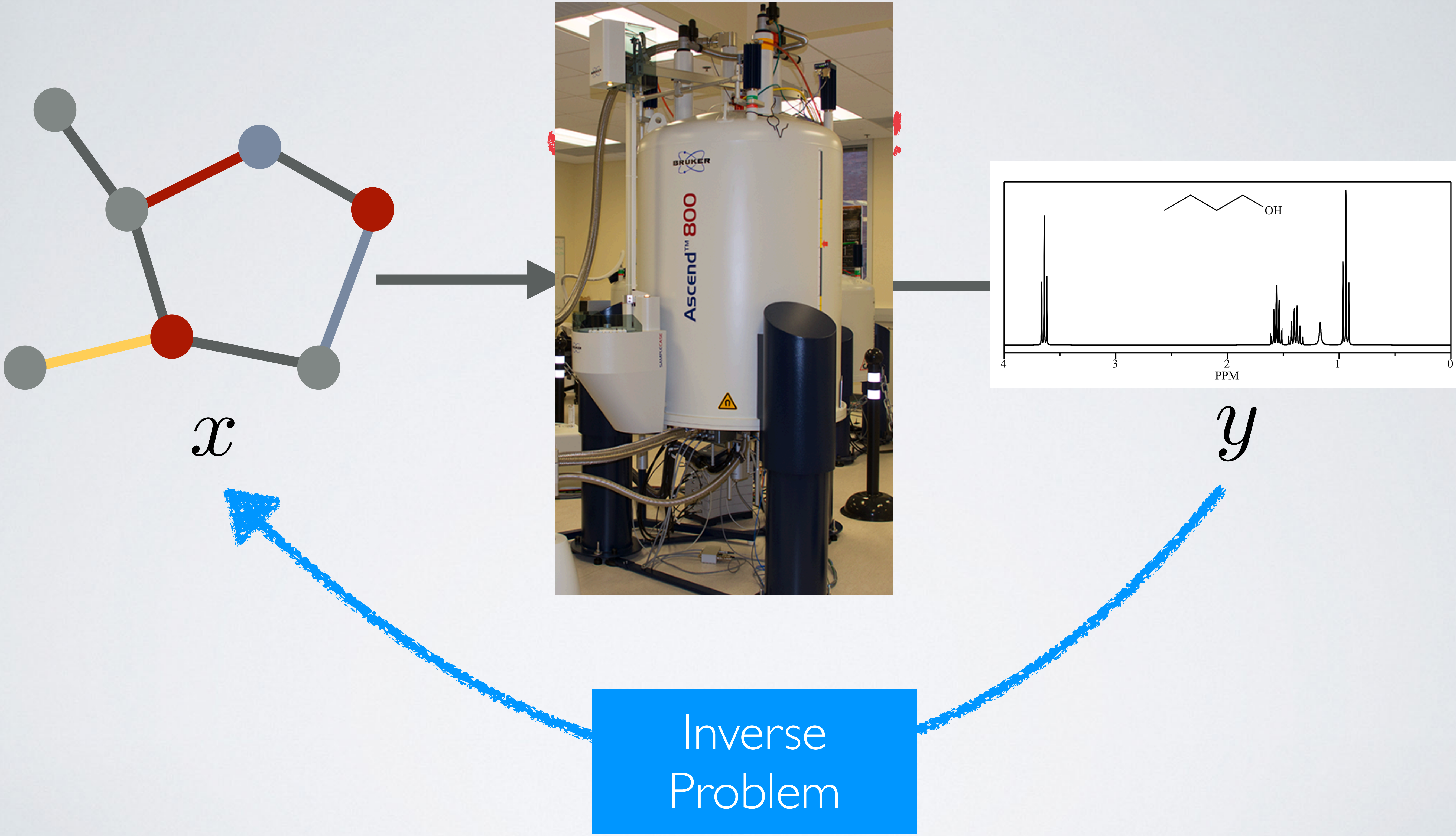


y

Inverse Problem

Spectroscopy as an

INVERSE PROBLEM



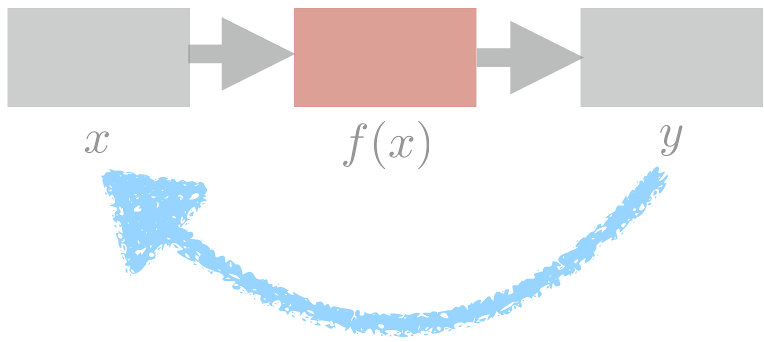
QUESTIONS FROM COMPUTER SCIENCE

How can we **recover graphs** from **measurements of their parts**?

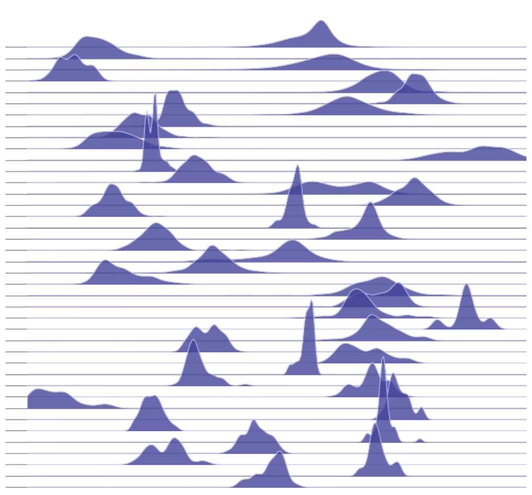
What is the best **representation of uncertainty** over graph structures?

What is the **optimal sequence of measurements** to identify a graph?

TODAY



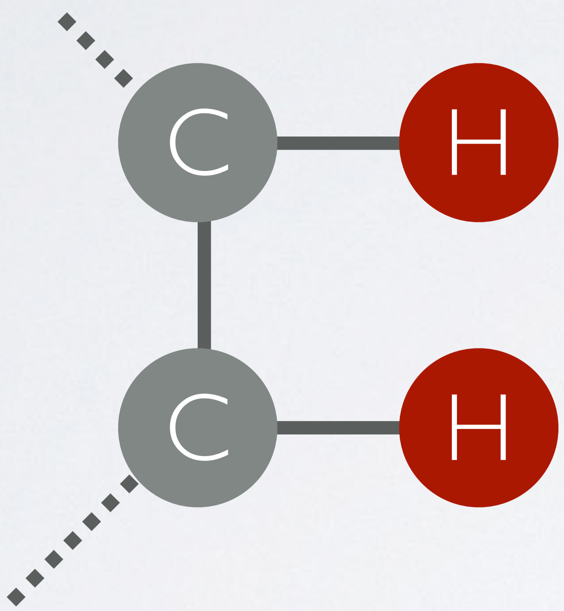
What are inverse problems?



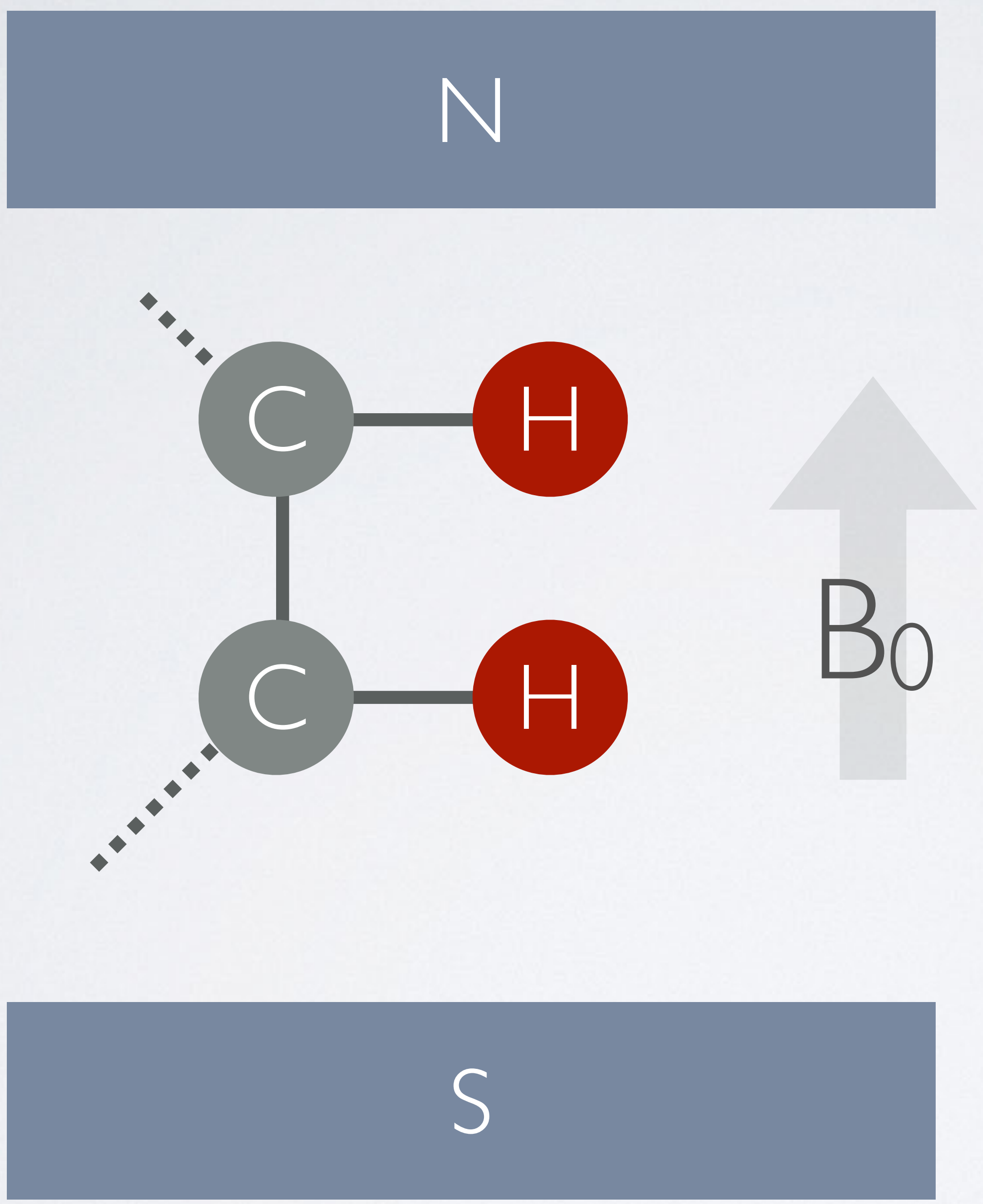
Spectroscopy:
The forward problem

Spectroscopy:
The inverse problem

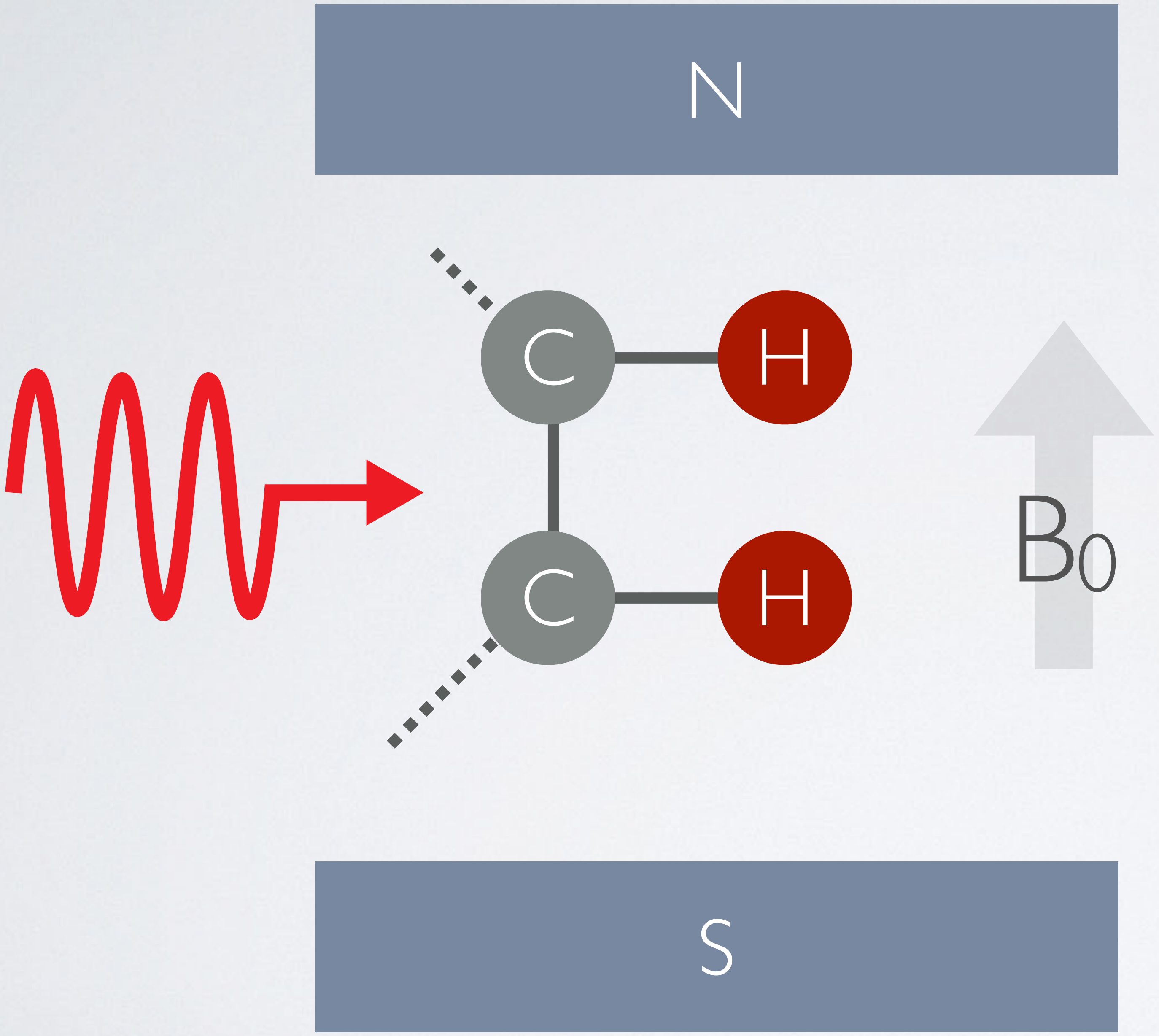
HOW DOES NMR WORK?



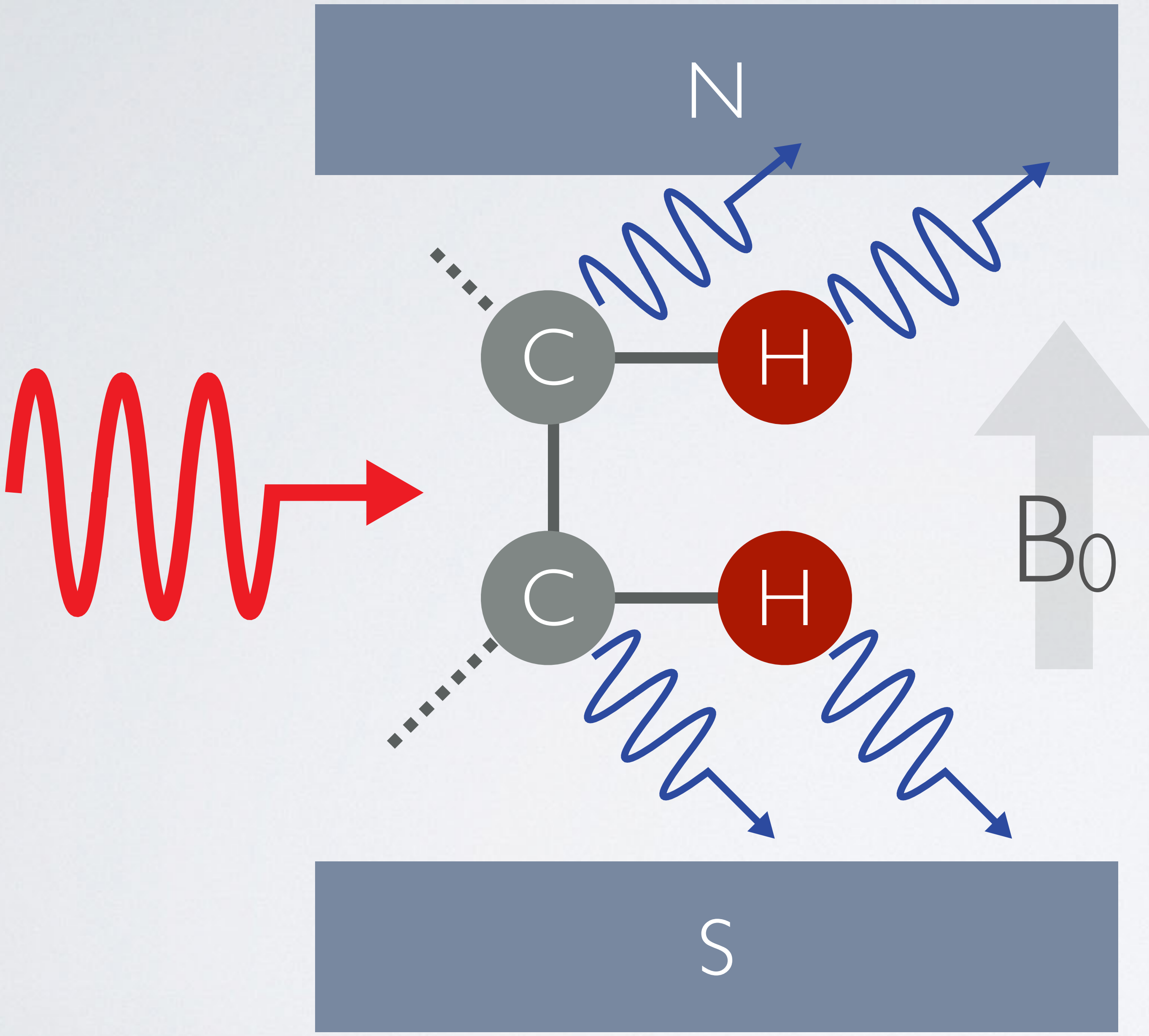
HOW DOES NMR WORK?



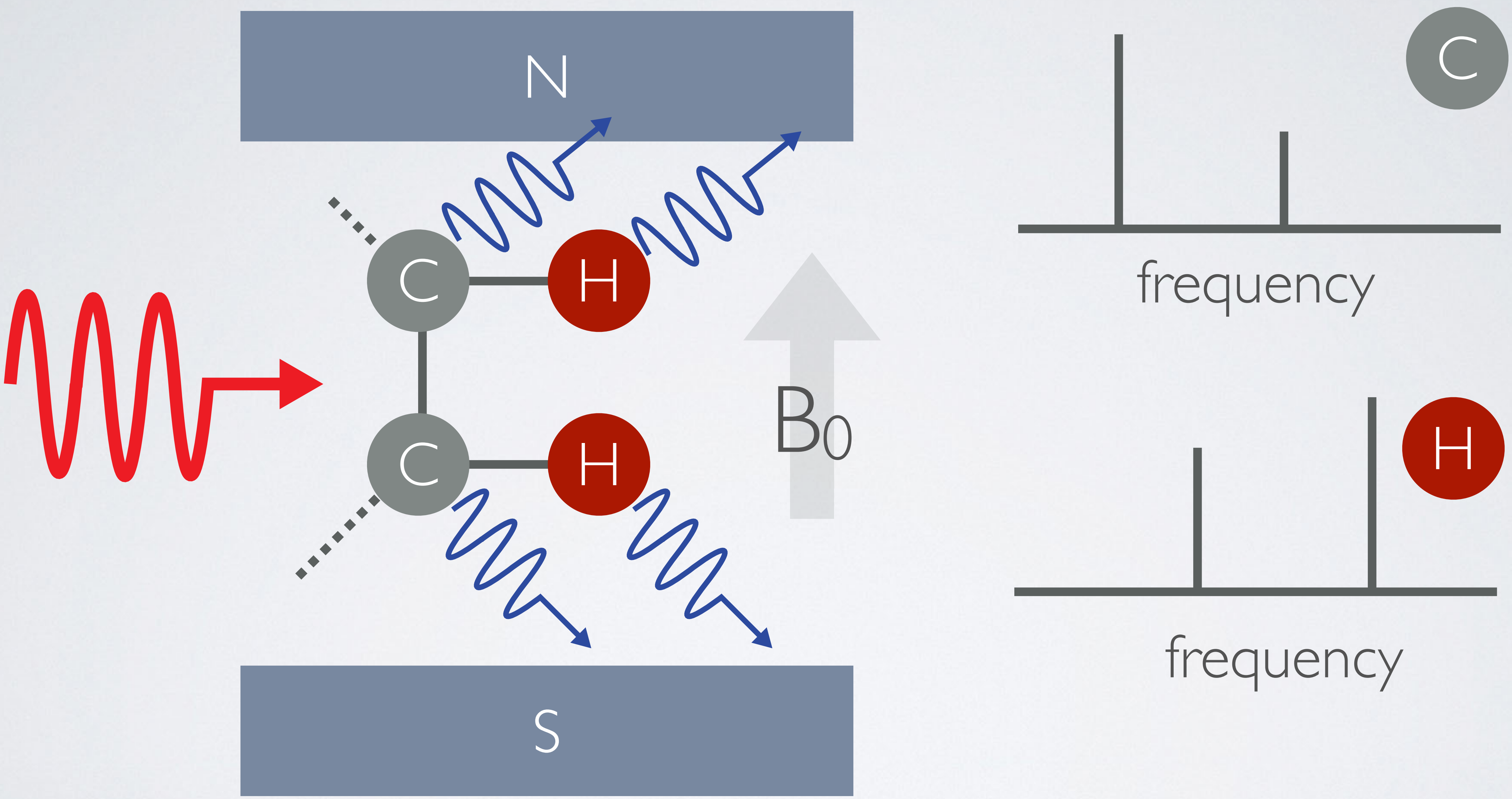
HOW DOES NMR WORK?

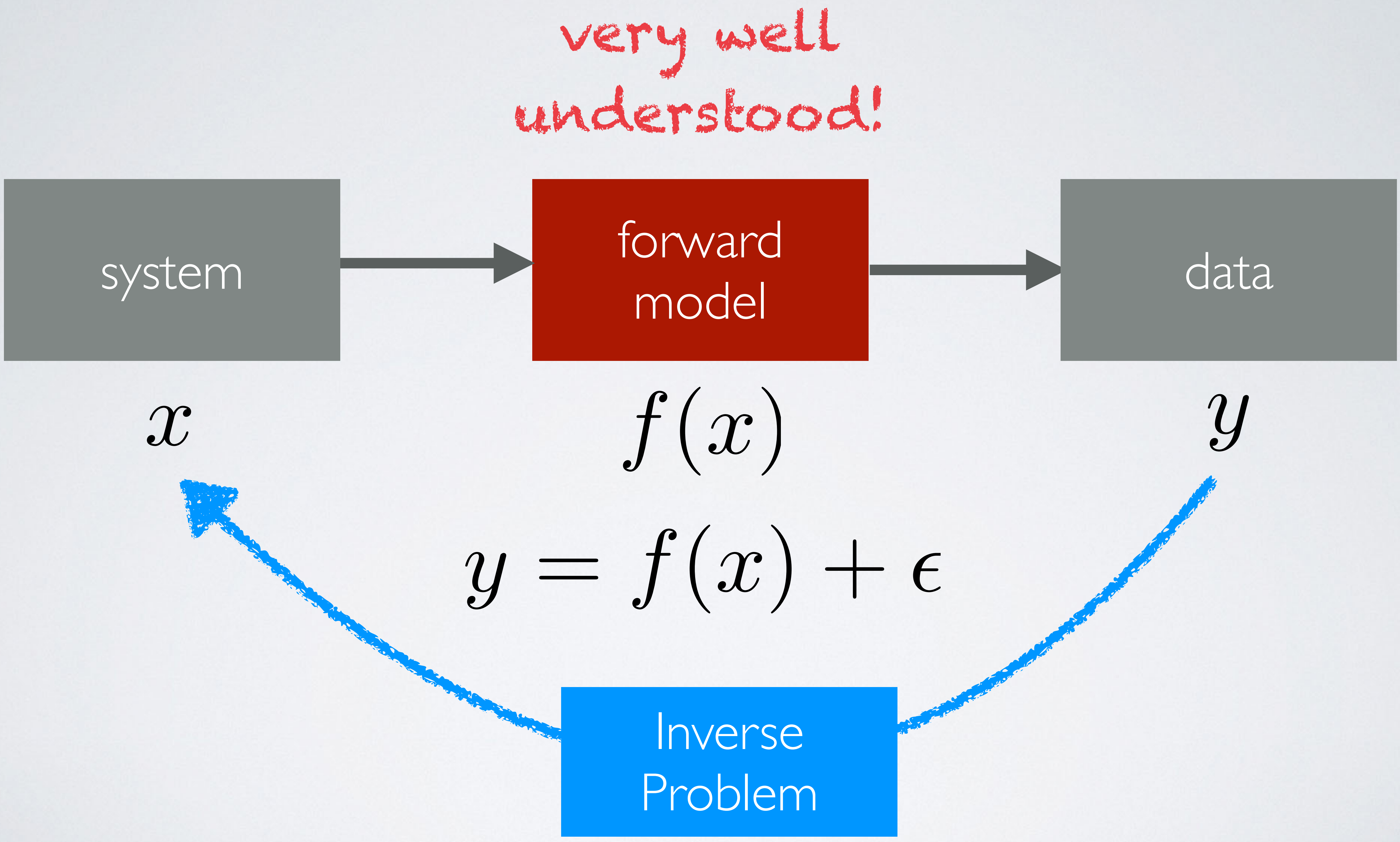


HOW DOES NMR WORK?



HOW DOES NMR WORK?




very well understood!
system
forward model
data
x
$f(x)$
y
$y = f(x) + \epsilon$
Inverse Problem

very well understood!

forward model

$$f(x)$$

very well understood!

Is it?

forward model

$$f(x)$$

FORWARD MODEL

Via computational chemistry

FORWARD MODEL

Via computational chemistry

Conformer Search

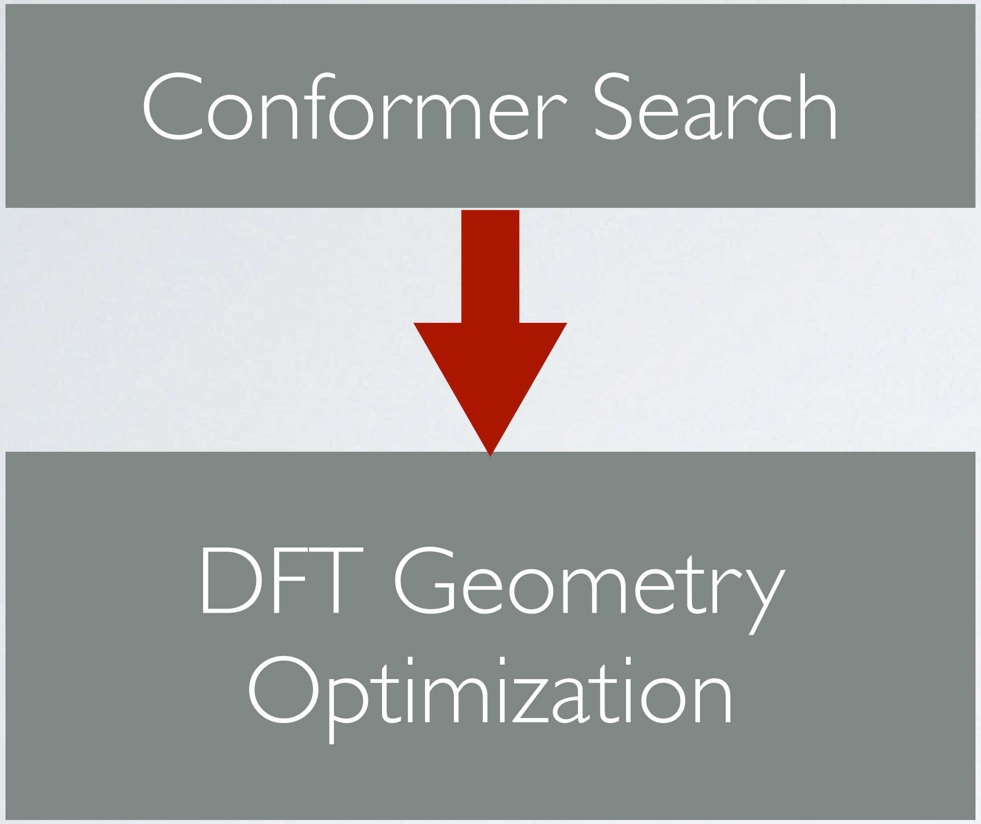
FORWARD MODEL

Via computational chemistry

Conformer Search

FORWARD MODEL

Via computational chemistry



FORWARD MODEL

Via computational chemistry

Conformer Search

DFT Geometry Optimization

FORWARD MODEL

Via computational chemistry

Conformer Search

↓

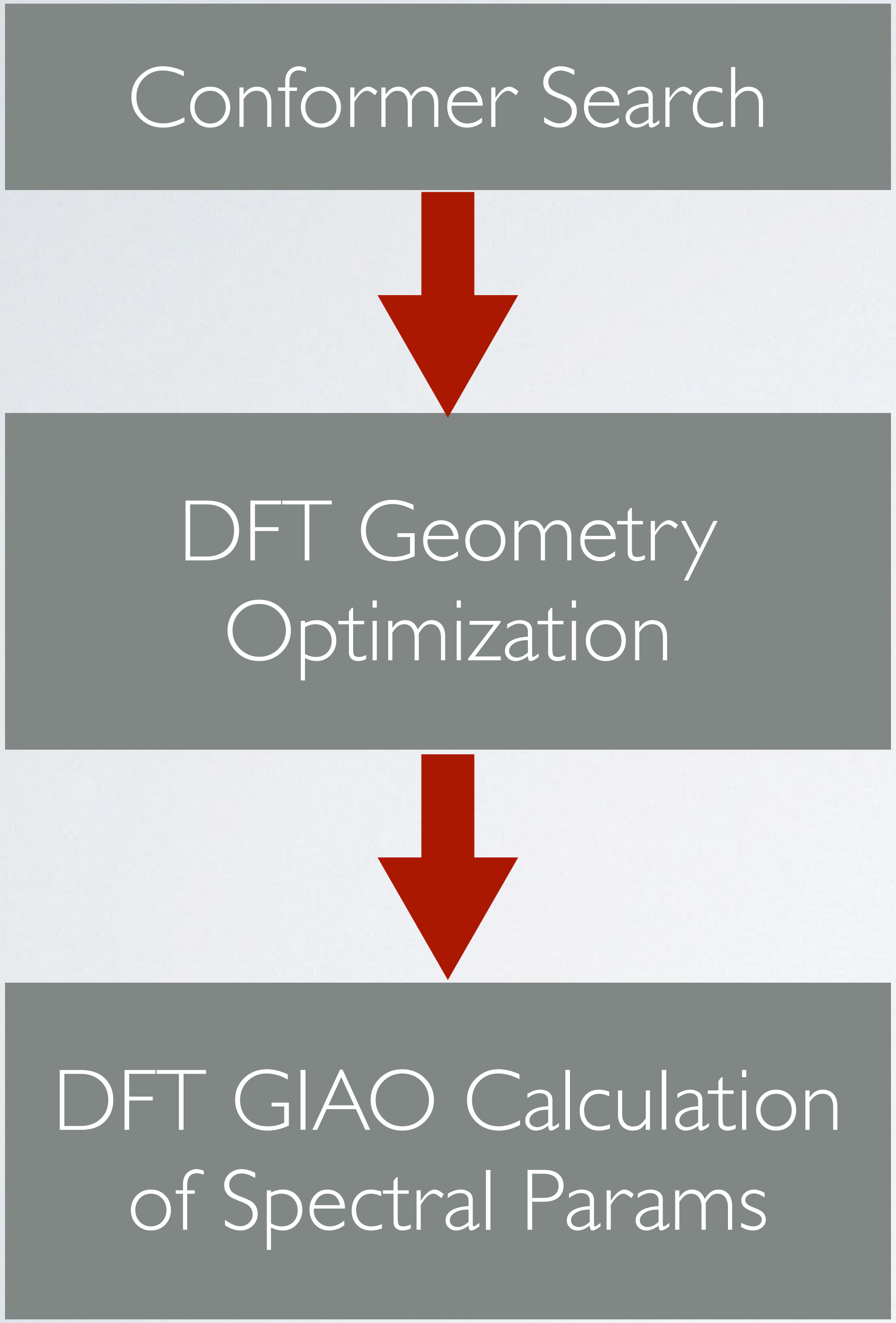
DFT Geometry Optimization

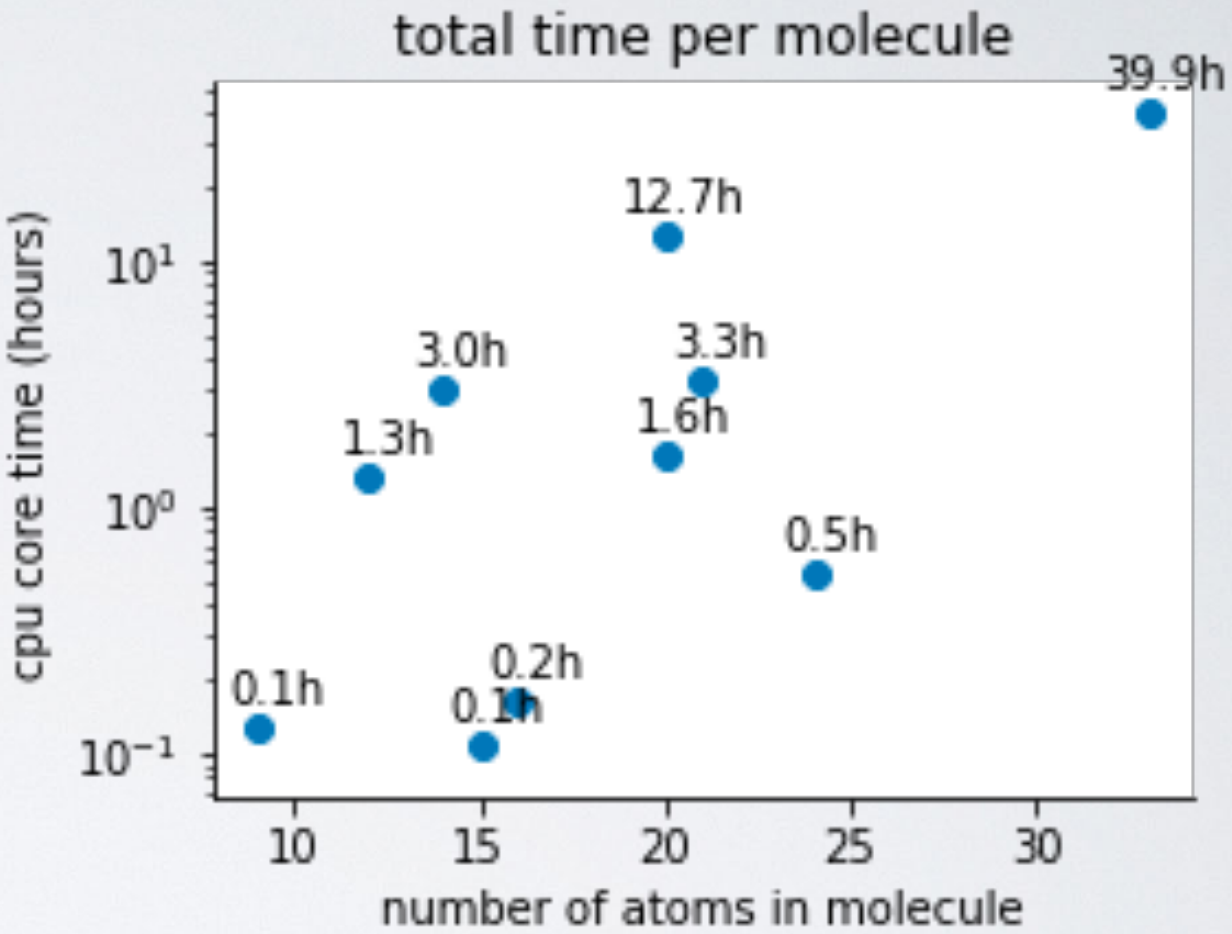
↓

DFT GIAO Calculation of Spectral Params

FORWARD MODEL

Via computational chemistry

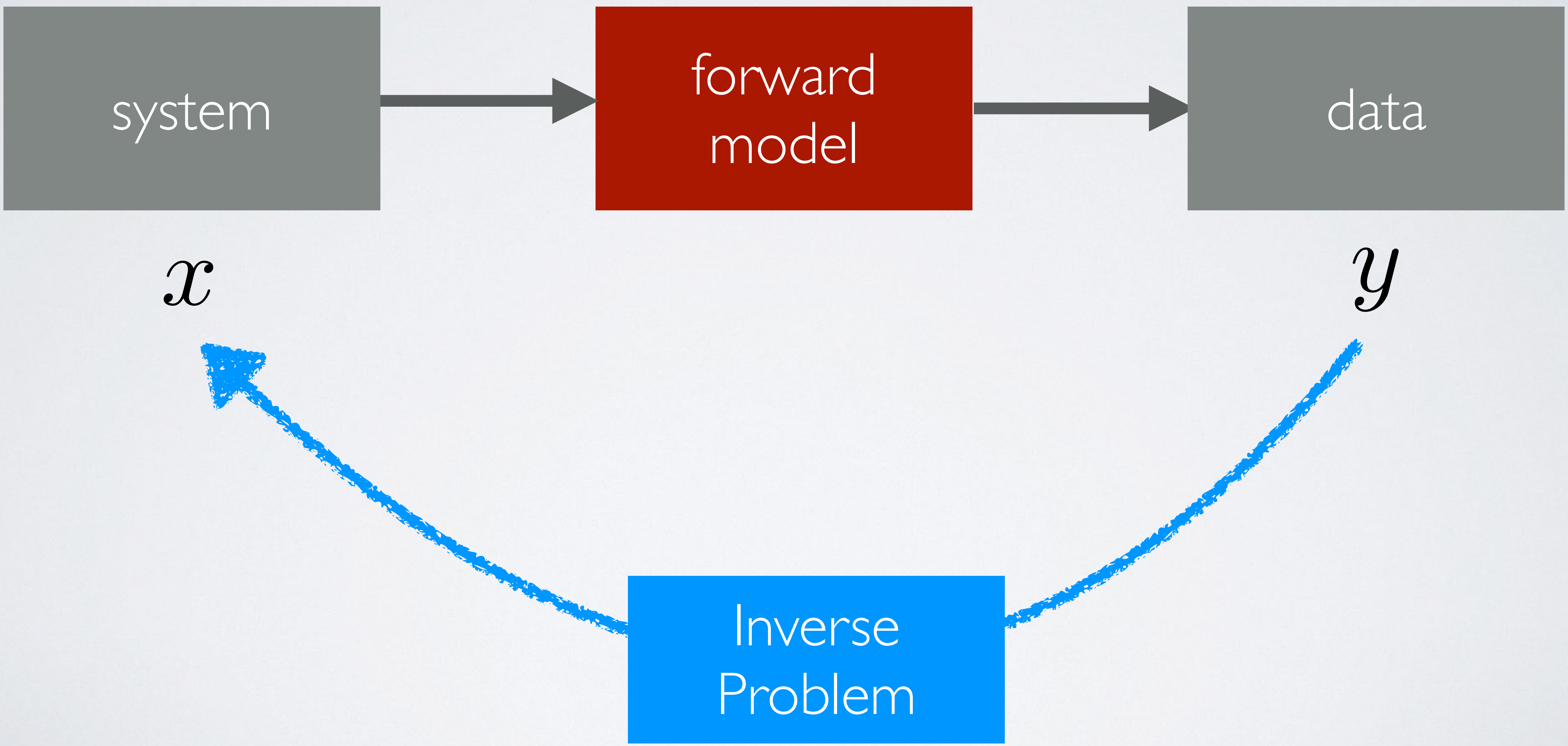




Ab initio is sloooow

SO OUR APPROACH

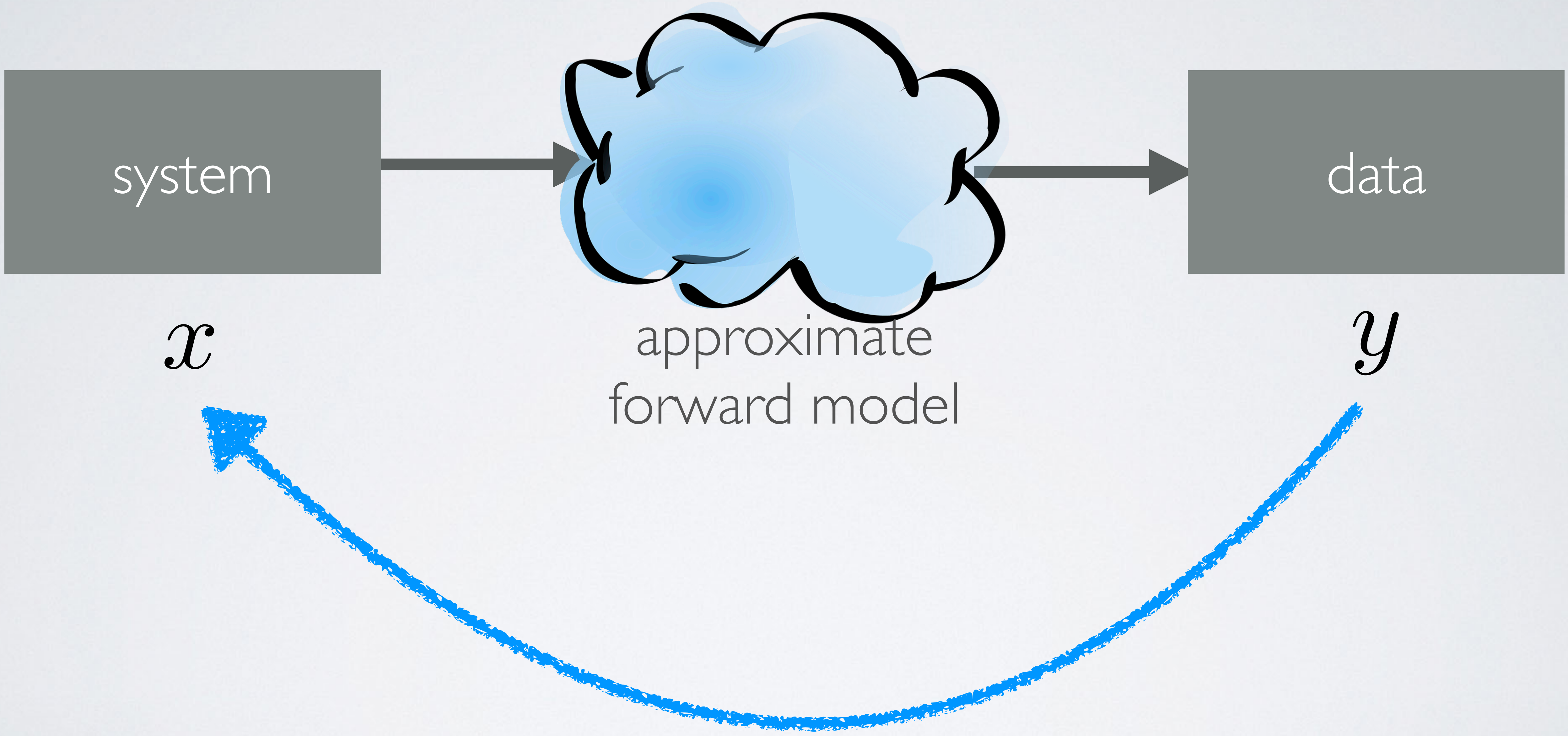
Jonas, E. & Kuhn, S. **Rapid prediction of NMR spectral properties with quantified uncertainty,** J Cheminform (2019) 11: 50



Jonas, Deep Imitation Learning for Molecular Inverse Problems, NeurIPS 2019

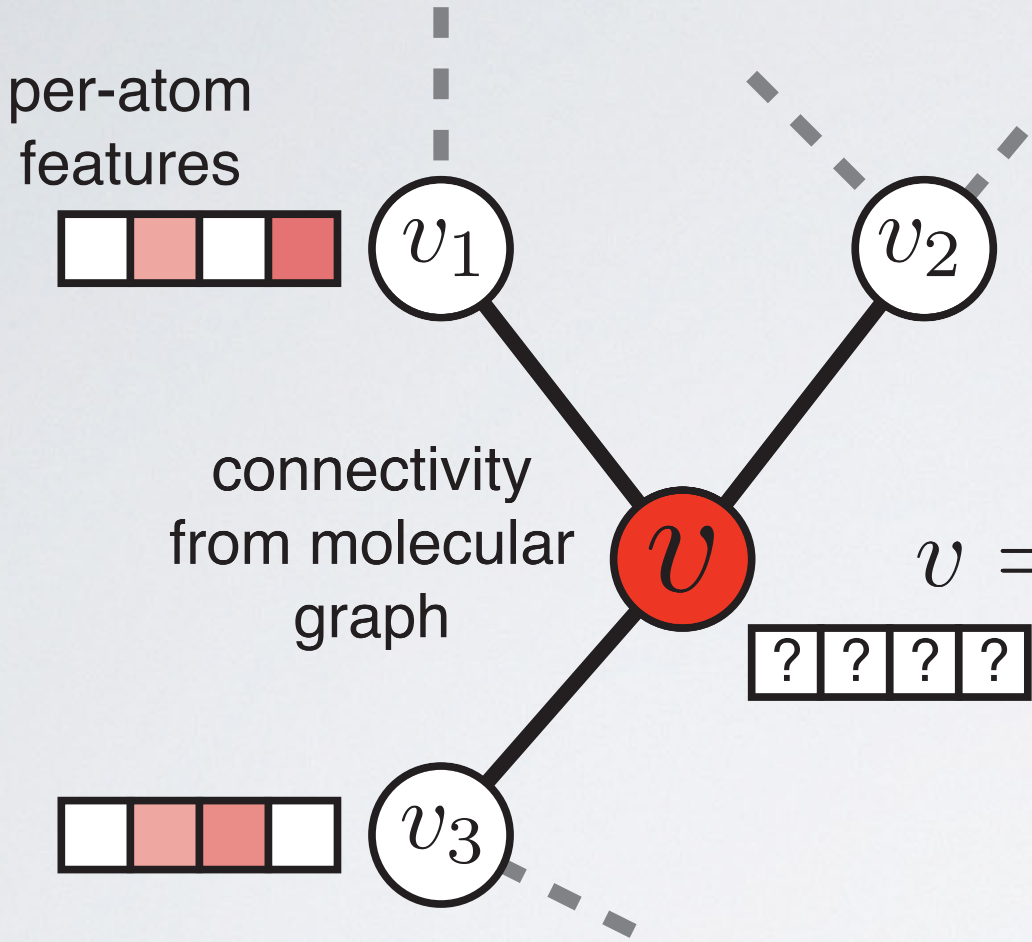
SO OUR APPROACH

Jonas, E. & Kuhn, S. **Rapid prediction of NMR spectral properties with quantified uncertainty,** J Cheminform (2019) 11: 50

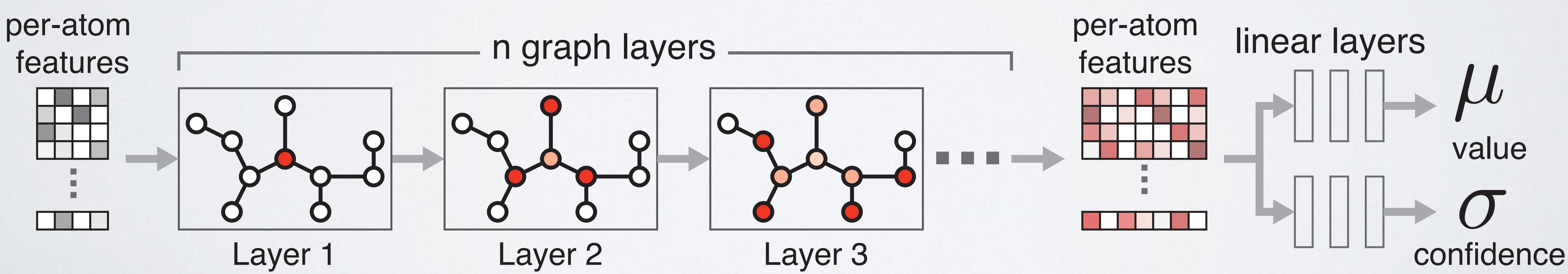


Jonas, Deep Imitation Learning for Molecular Inverse Problems, NeurIPS 2019

GRAPH NEURAL NETWORKS



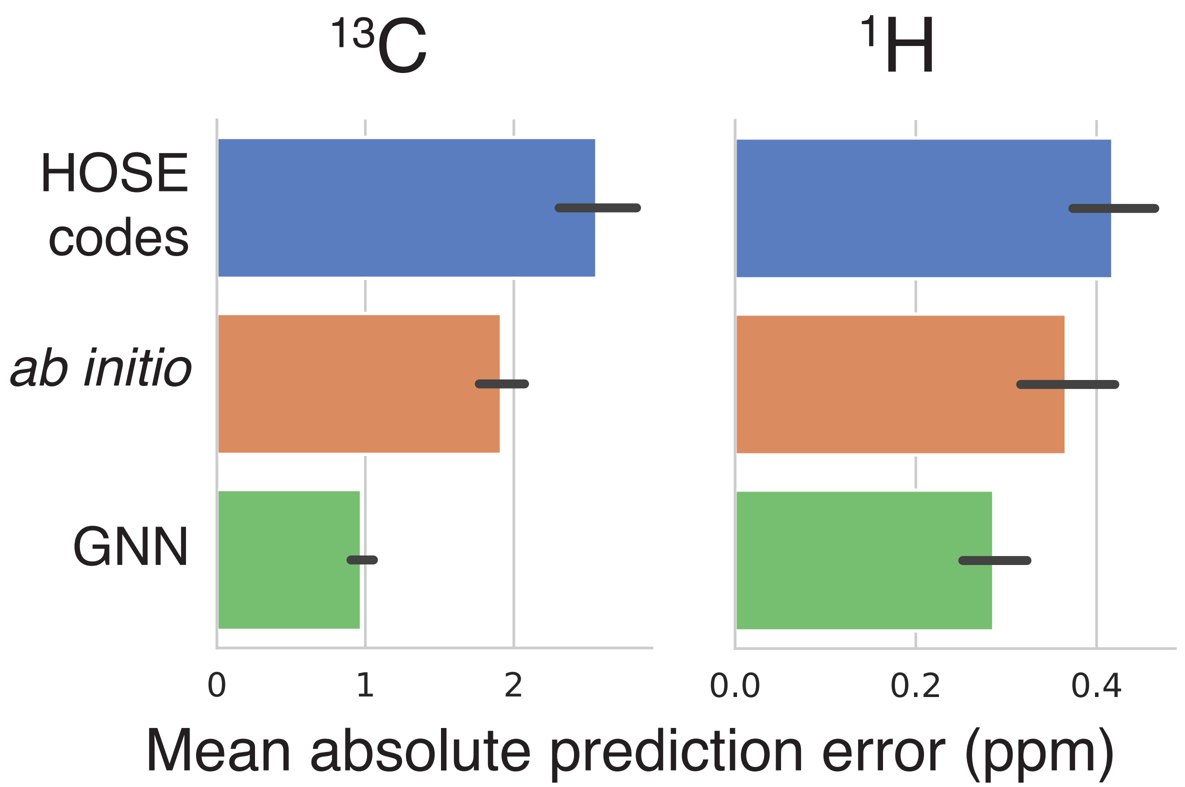
$$v = f\left(\sum_{i \in \mathcal{N}(v)} w^T v_i\right)$$



COMPARISON

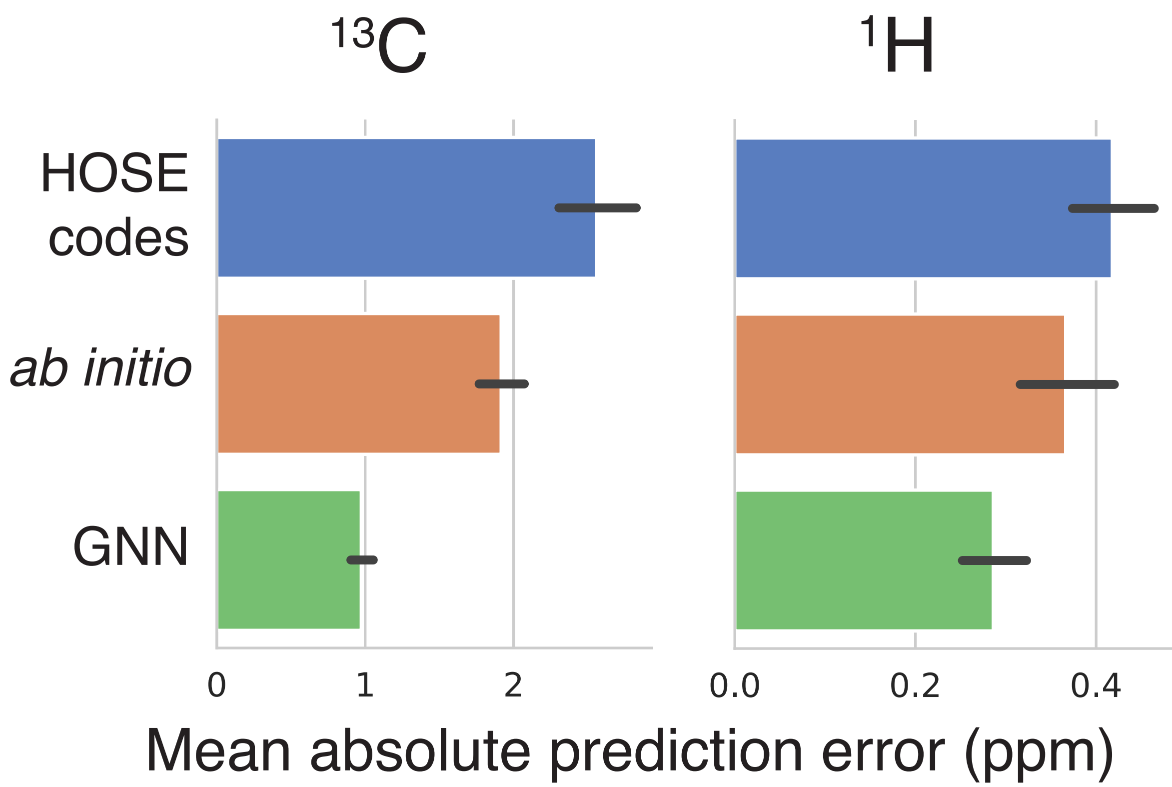
(all error bars are bootstrap 95% confidence interval of mean)

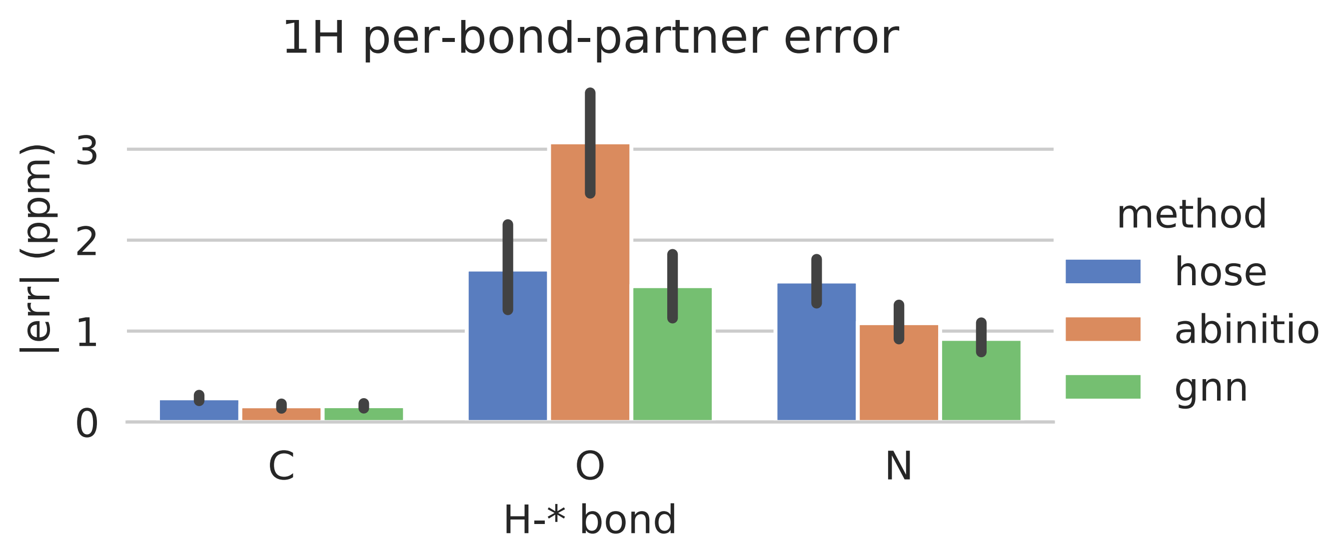
COMPARISON



(all error bars are bootstrap 95% confidence interval of mean)

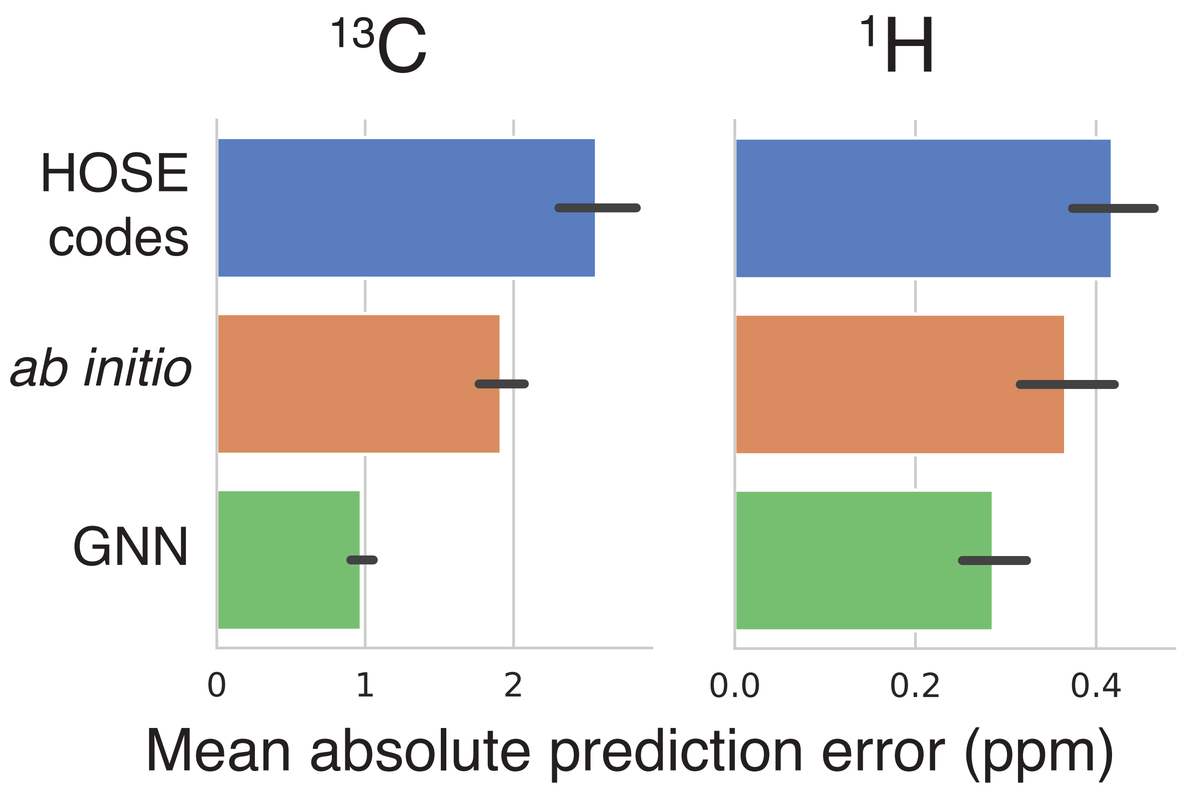
COMPARISON

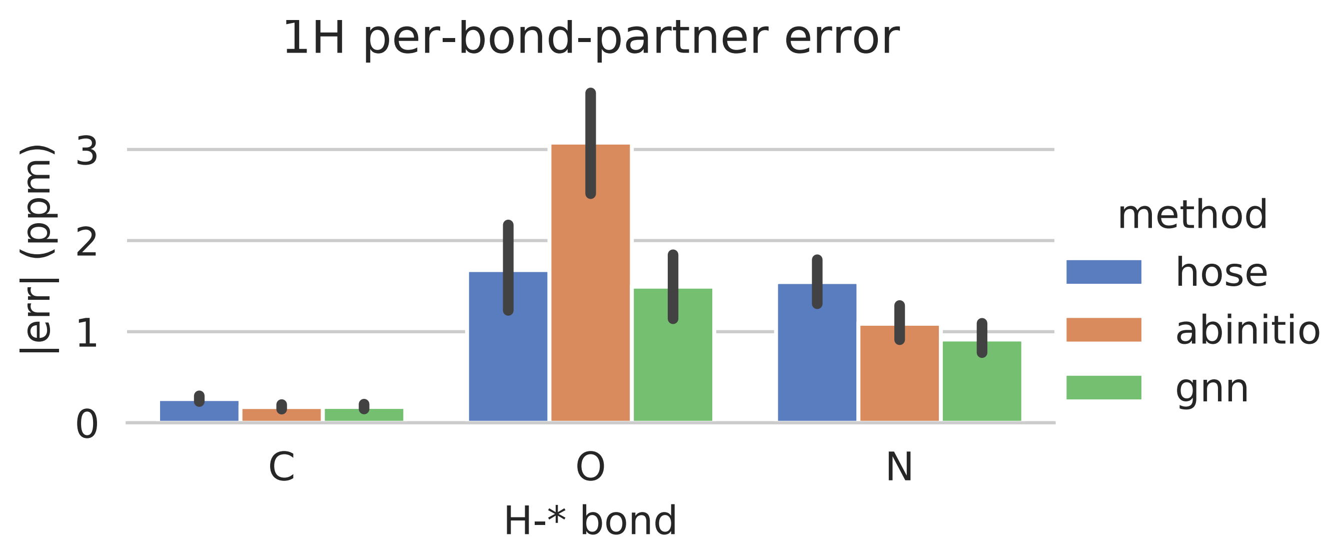


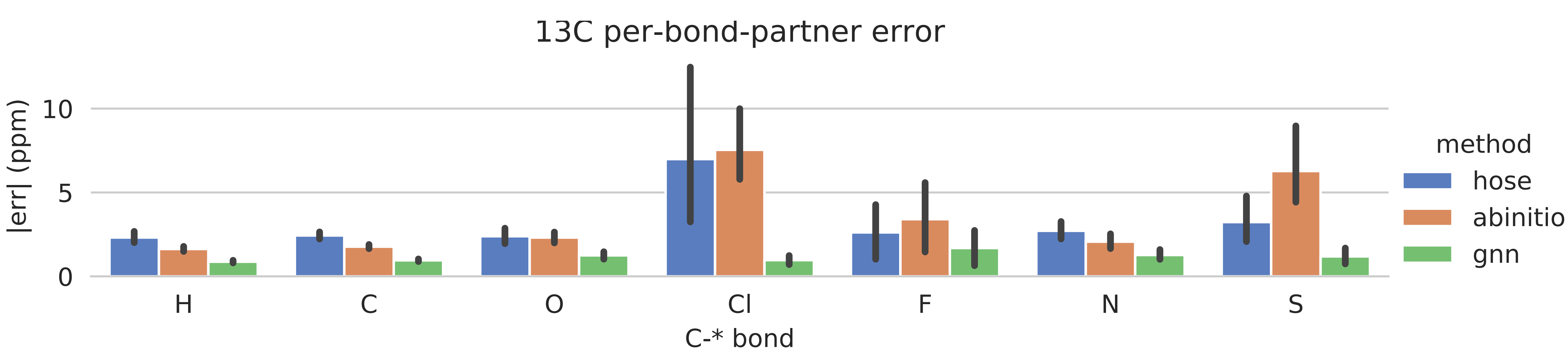


(all error bars are bootstrap 95% confidence interval of mean)

COMPARISON







(all error bars are bootstrap 95% confidence interval of mean)

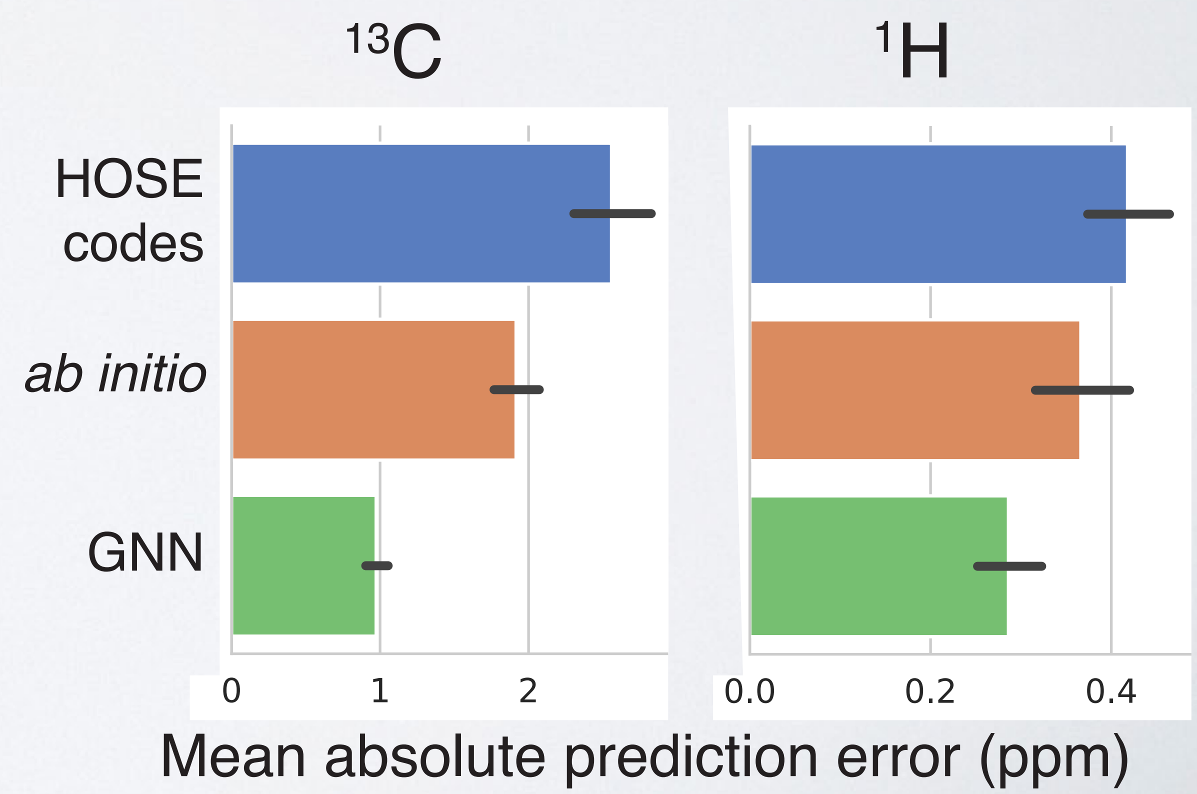
FAST AND ACCURATE

FAST AND ACCURATE

Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec

FAST AND ACCURATE

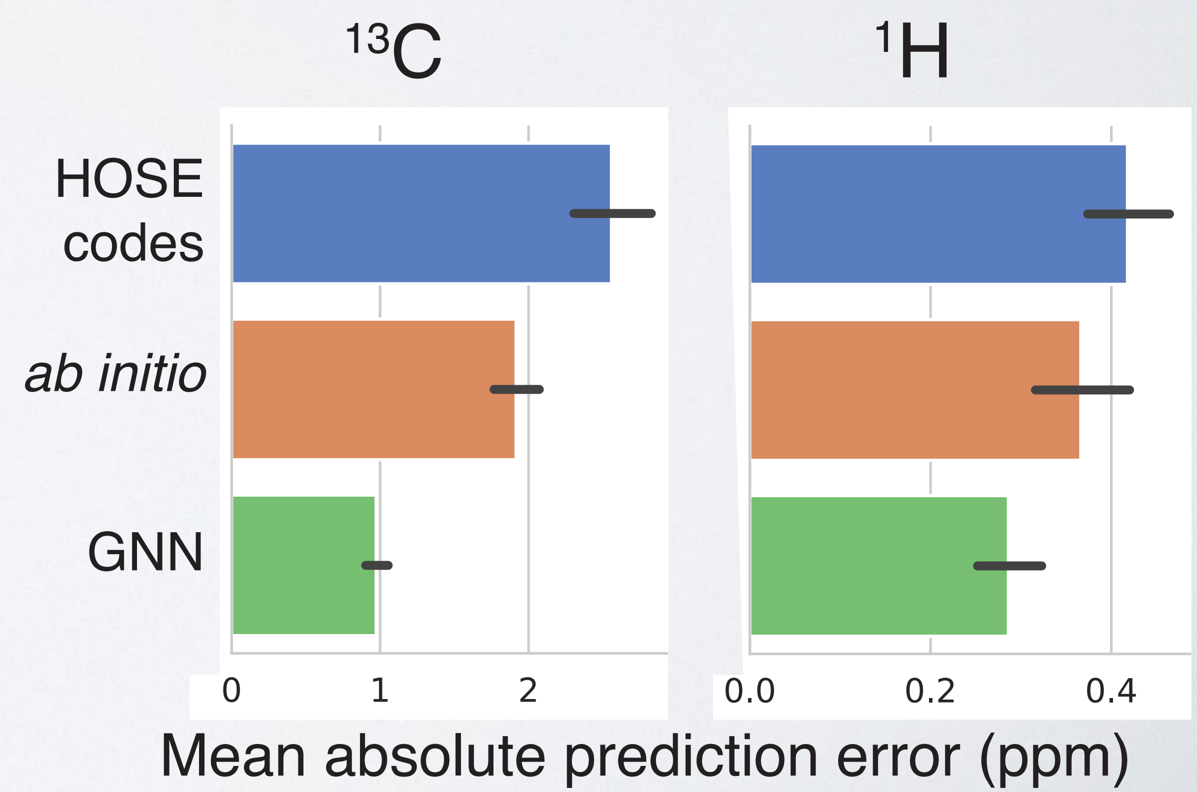
Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec



FAST AND ACCURATE

Ongoing work

Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec

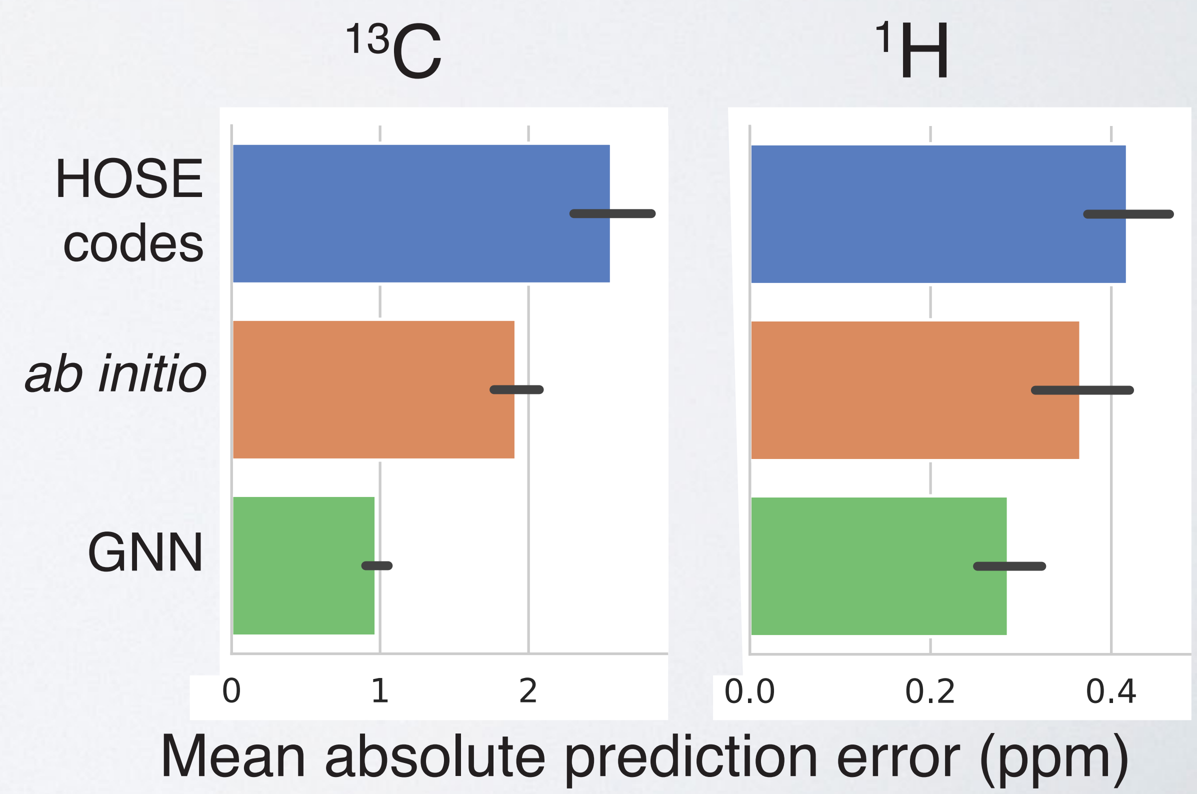


FAST AND ACCURATE

Ongoing work

- Incorporating richer molecular structure / space

Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec

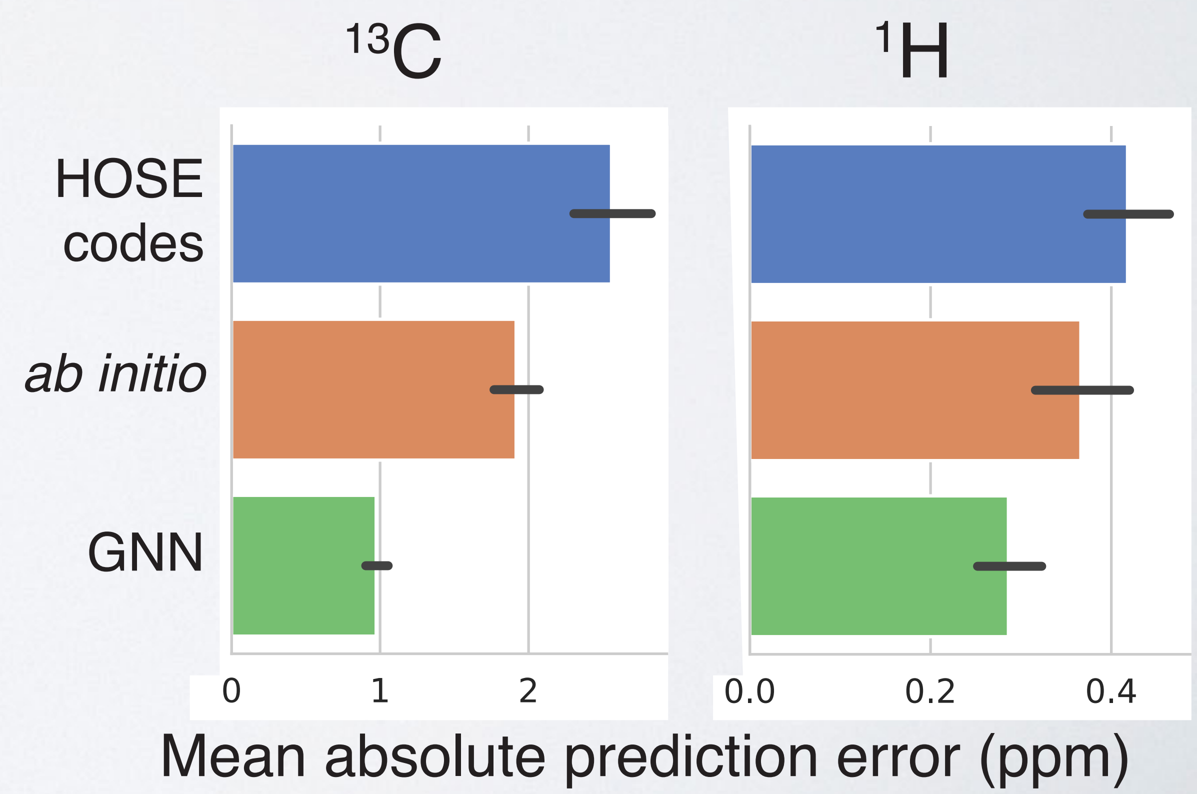


FAST AND ACCURATE

Ongoing work

- Incorporating richer molecular structure / space
- Predict scalar coupling constants as well (very useful)

Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec

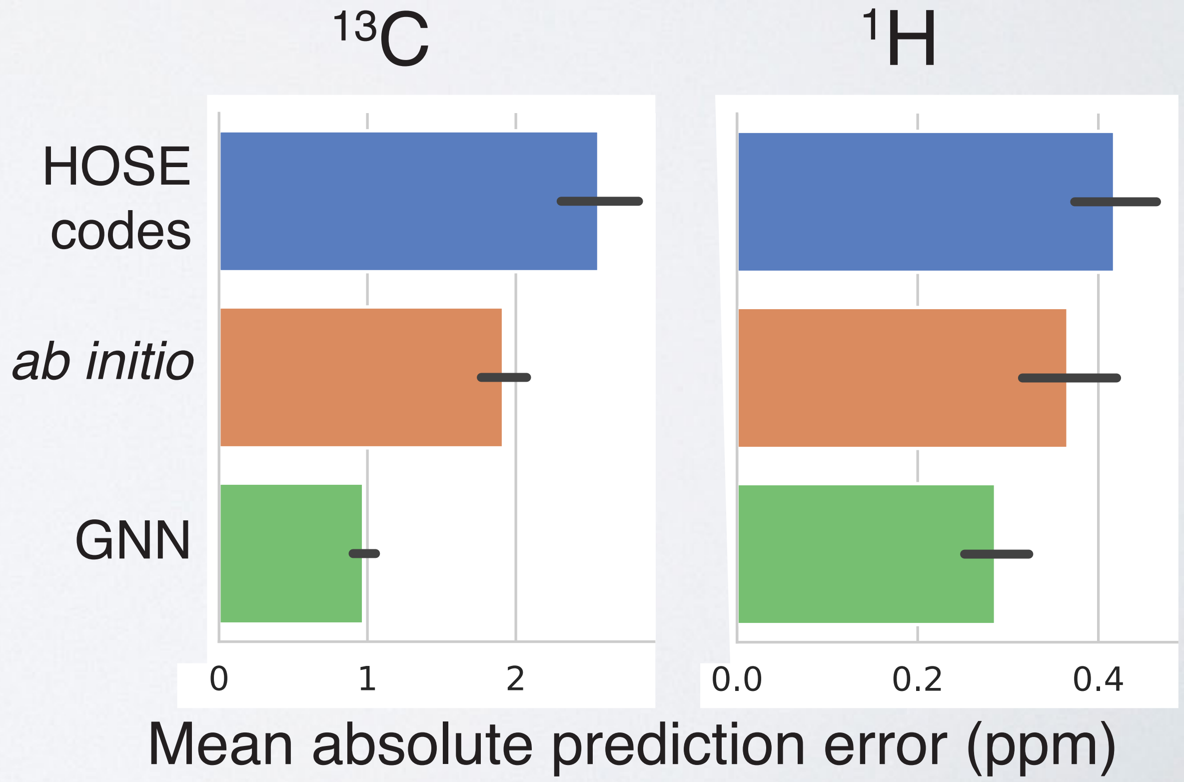


FAST AND ACCURATE

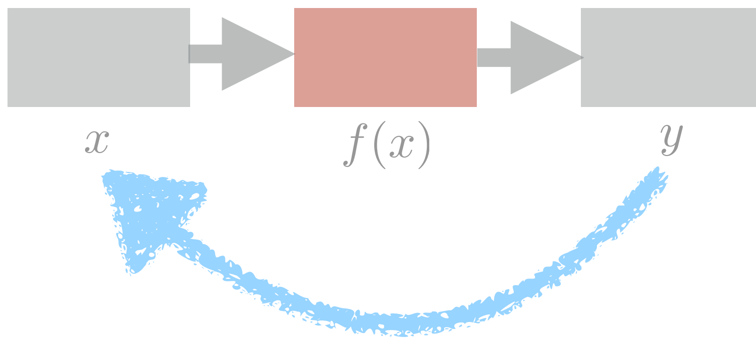
Ongoing work

- Incorporating richer molecular structure / space
- Predict scalar coupling constants as well (very useful)
- Active learning with ab initio data

Method	note	per-mol	per-nucleus
GNN[1]		3.6 ms	56 μs
HOSE[2]	13C	29 ms	2 ms
	1H	34 ms	4 ms
DFT[3]	geom. opt	556 sec	36 sec
	GIAO	256 sec	16 sec

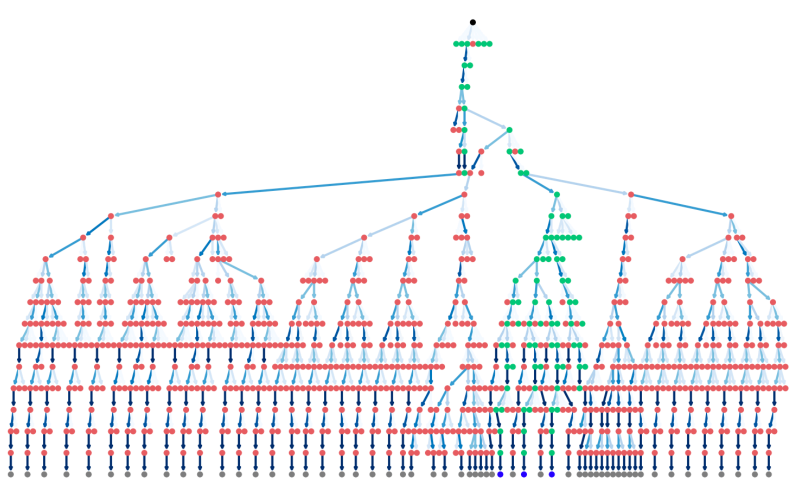


TODAY



What are
inverse problems?

Spectroscopy:
The forward problem

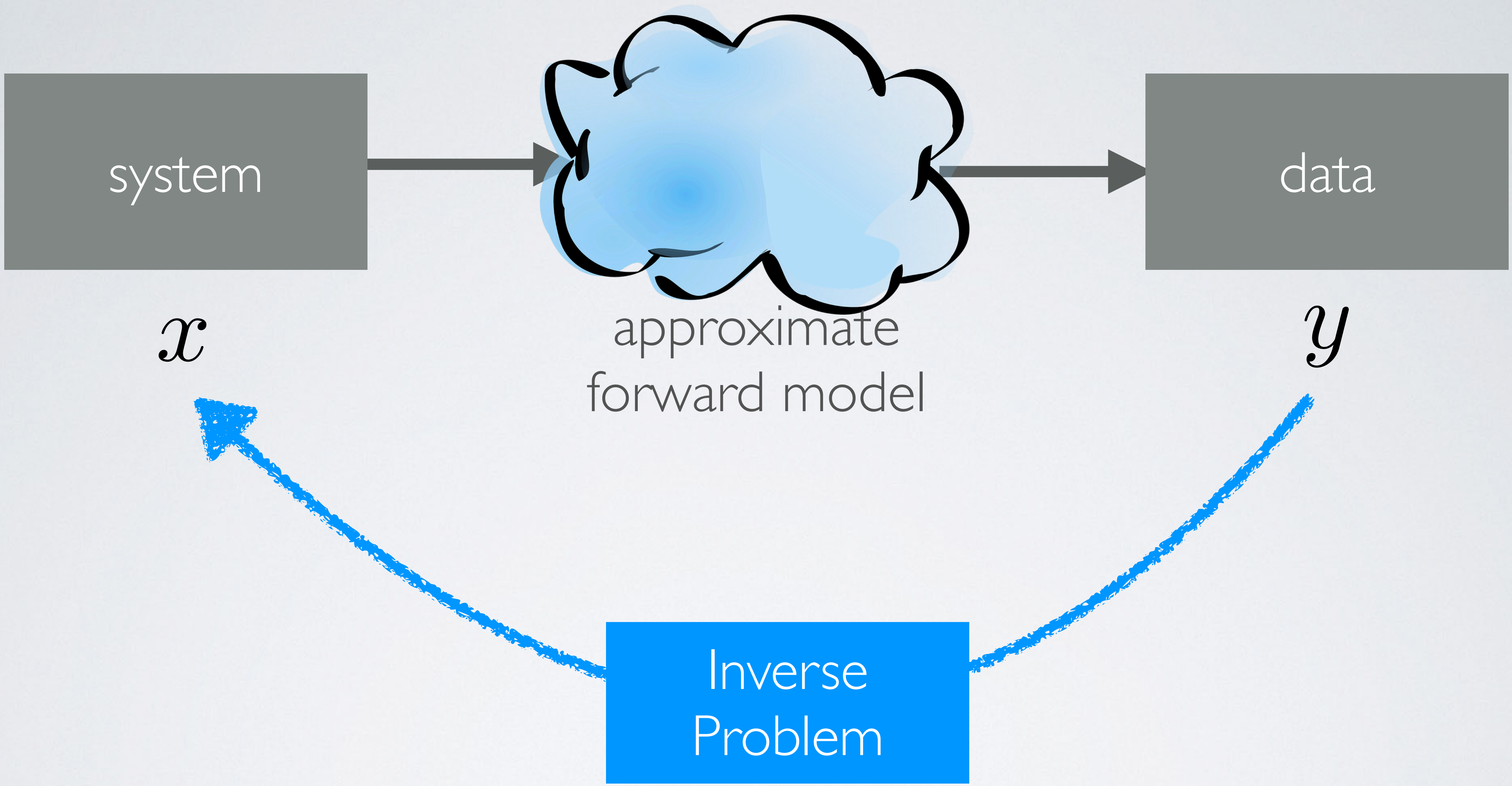


Spectroscopy:
The inverse problem

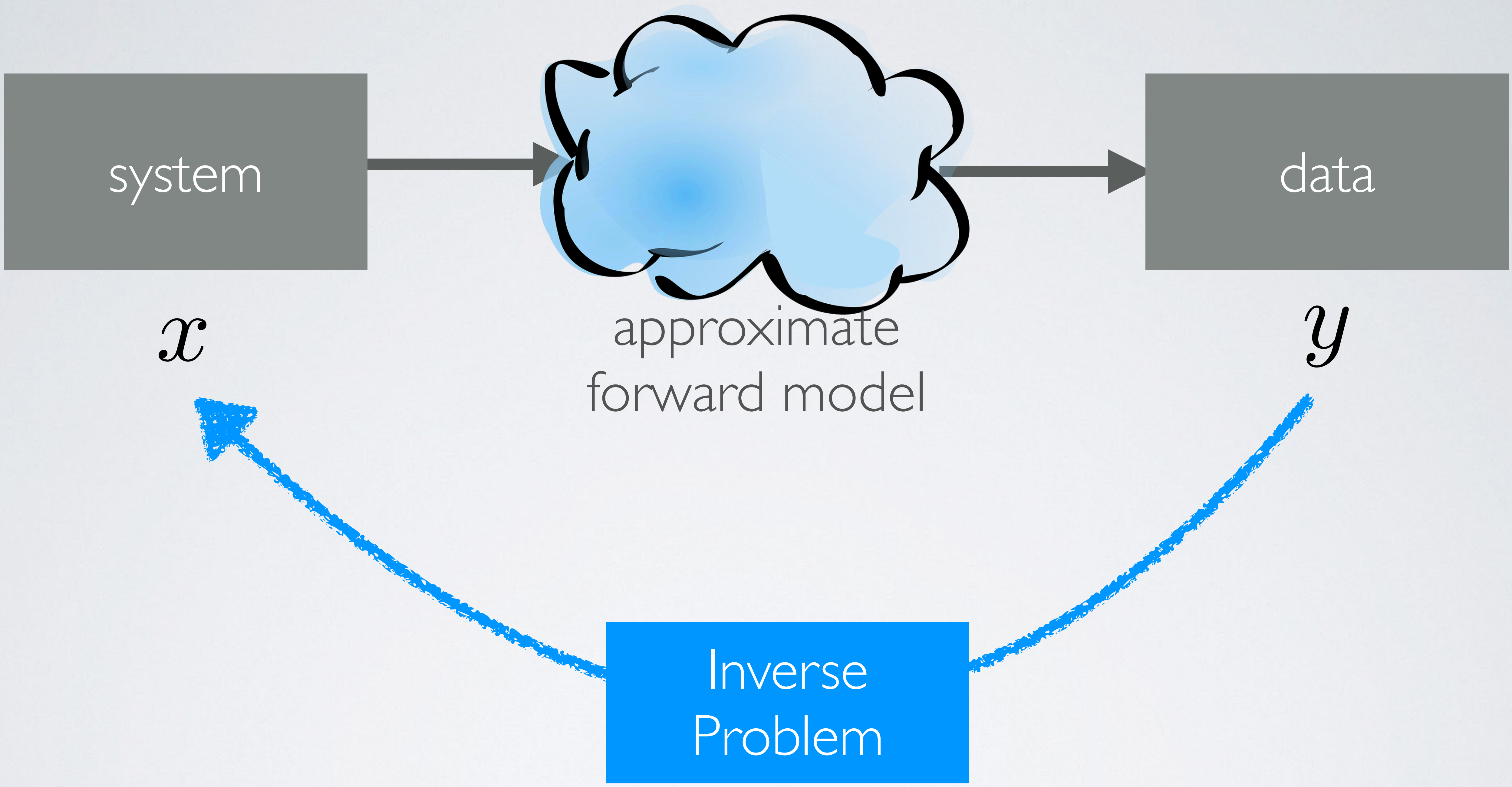
Jonas, E. & Kuhn, S. **Rapid prediction of NMR spectral properties with quantified uncertainty,** J Cheminform (2019) 11:50

x y

Jonas, E. & Kuhn, S. **Rapid prediction of NMR spectral properties with quantified uncertainty,** J Cheminform (2019) 11: 50



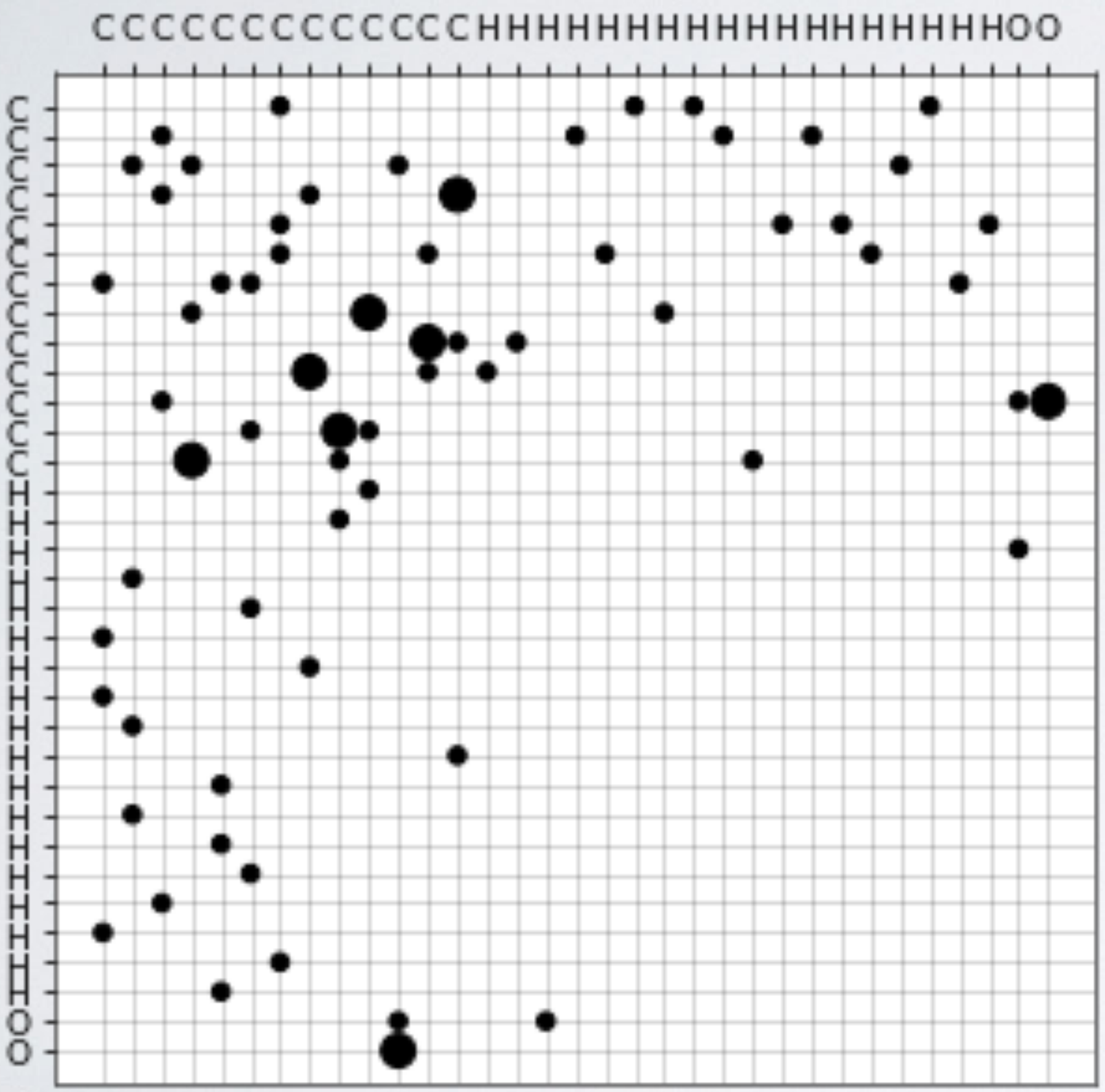
Jonas, E. & Kuhn, S. **Rapid prediction of NMR spectral properties with quantified uncertainty,** J Cheminform (2019) 11:50

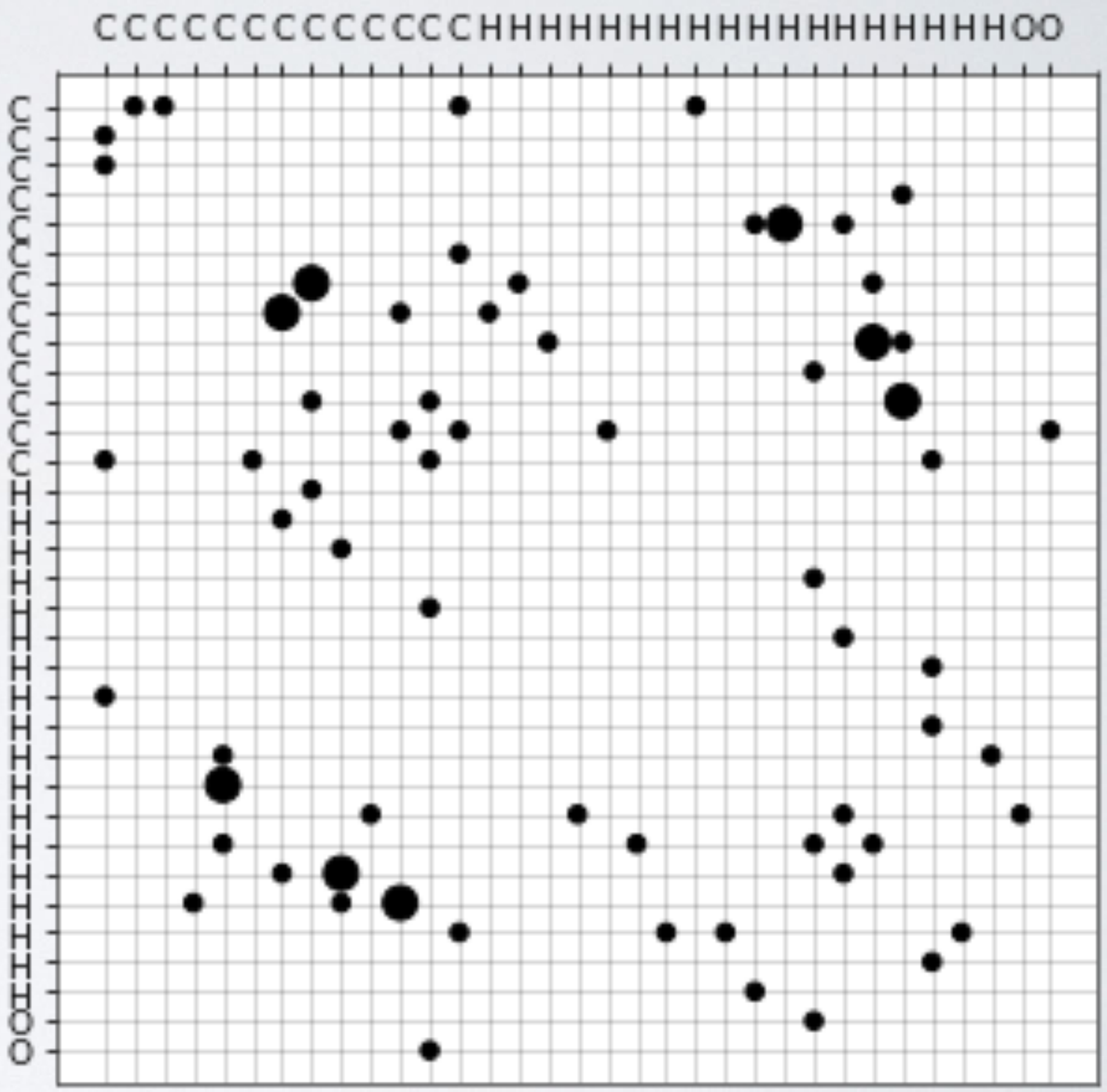


Jonas, **Deep Imitation Learning for Molecular Inverse Problems** (NeurIPS 2019)

LET'S TALK ISOMORPHISM

Are these two graphs the same?





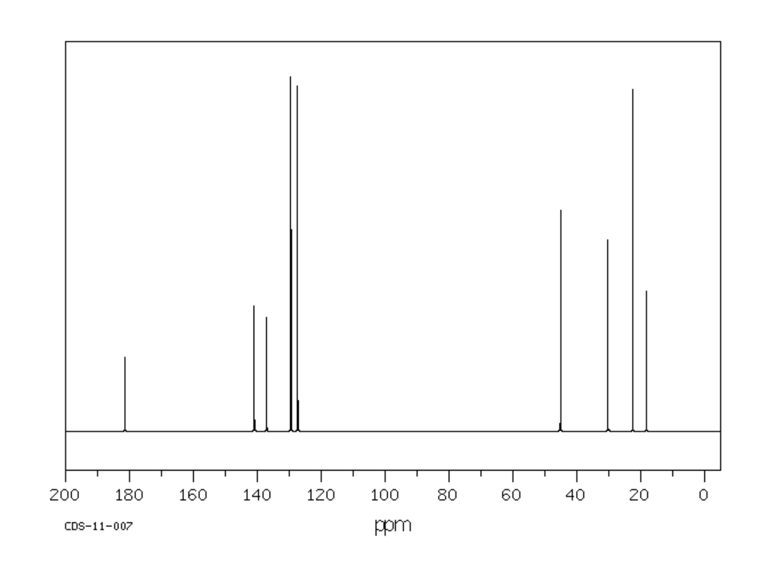
PROBLEM FORMULATION

PROBLEM FORMULATION

Molecular formula

$C_{13}H_{18}O_2$

Spectrum

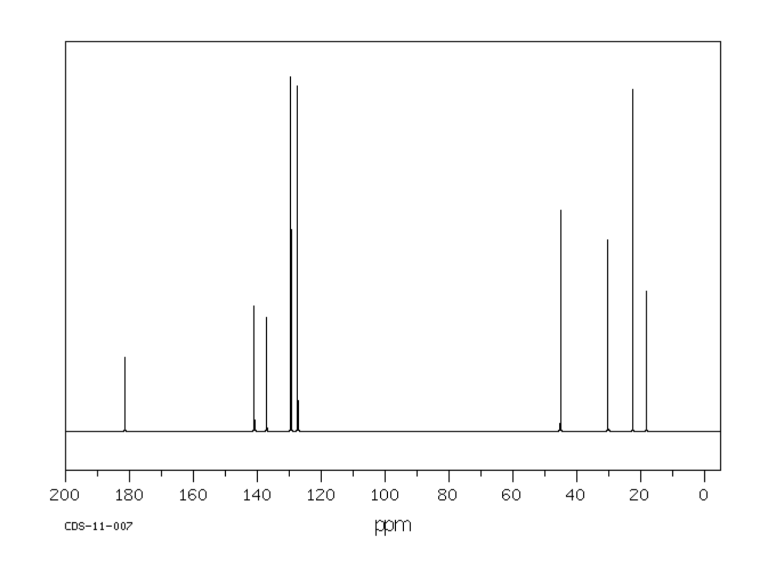


PROBLEM FORMULATION

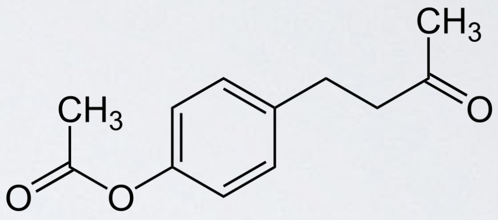
Molecular formula

$C_{13}H_{18}O_2$

Spectrum



Structure

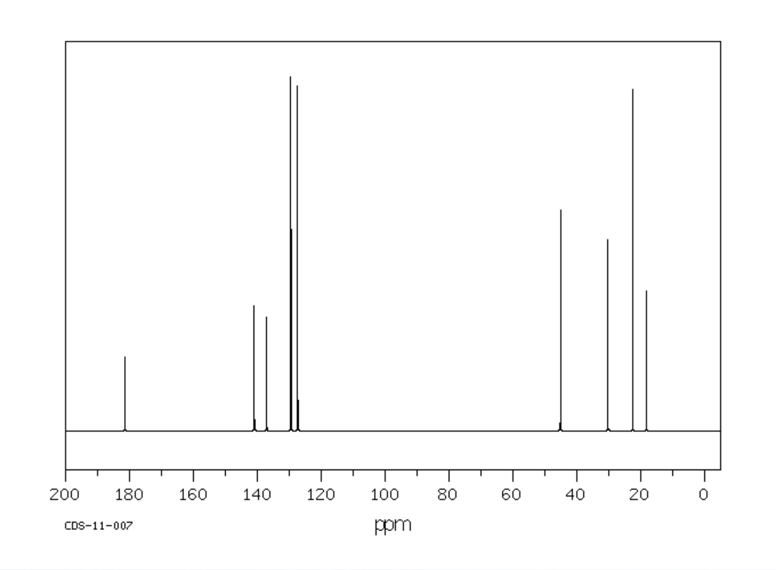


PROBLEM FORMULATION

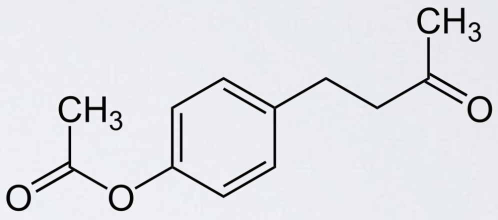
Molecular formula

$C_{13}H_{18}O_2$

Spectrum



Structure



vertices

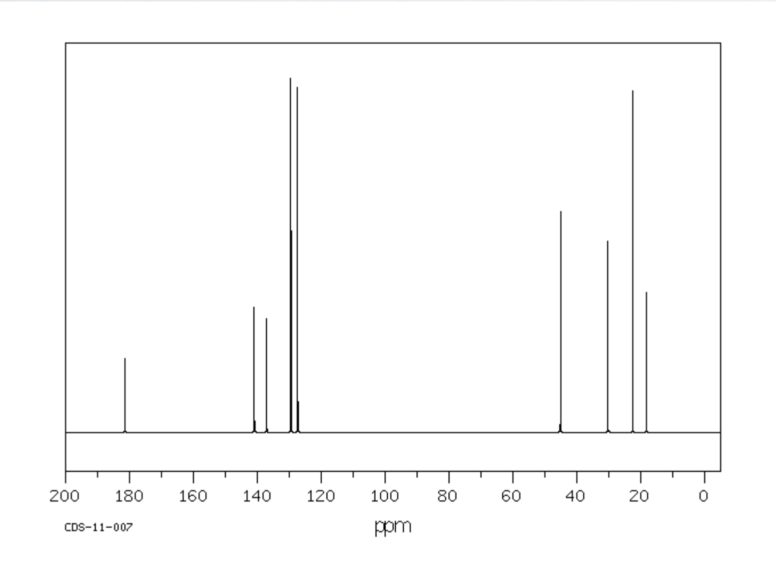
per-vertex properties

PROBLEM FORMULATION

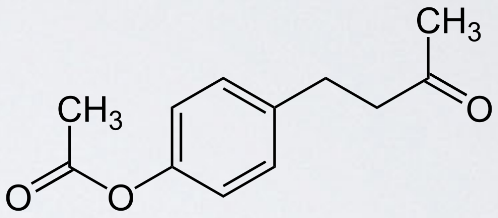
Molecular formula

$C_{13}H_{18}O_2$

Spectrum



Structure



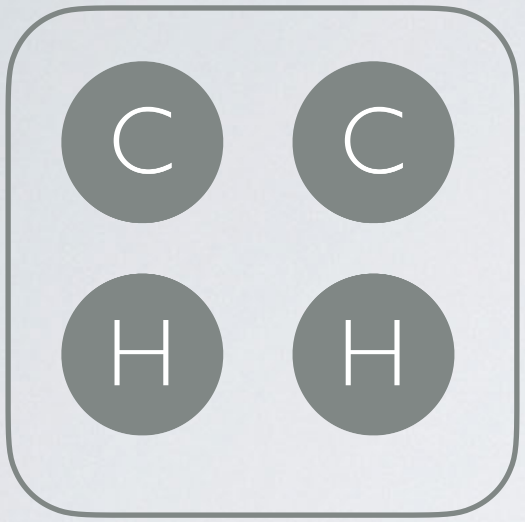
vertices

per-vertex properties

edges

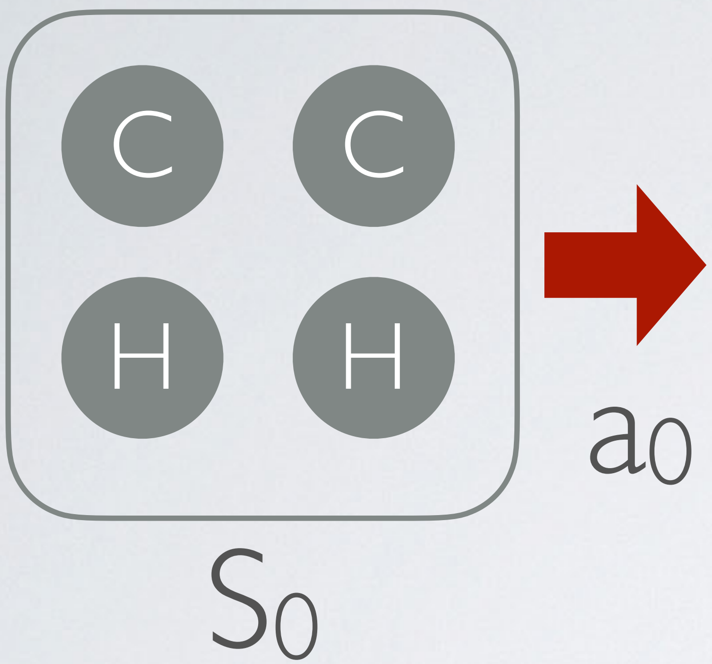
SEQUENTIAL BOND FORMATION

SEQUENTIAL BOND FORMATION

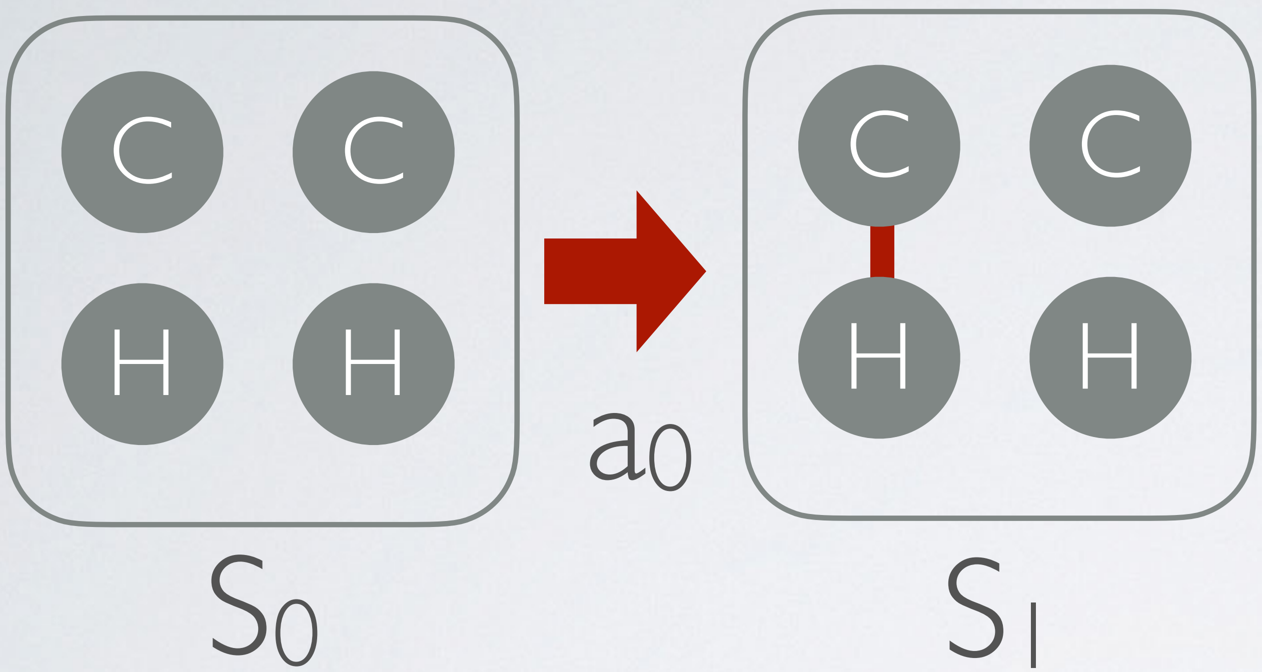


S_0

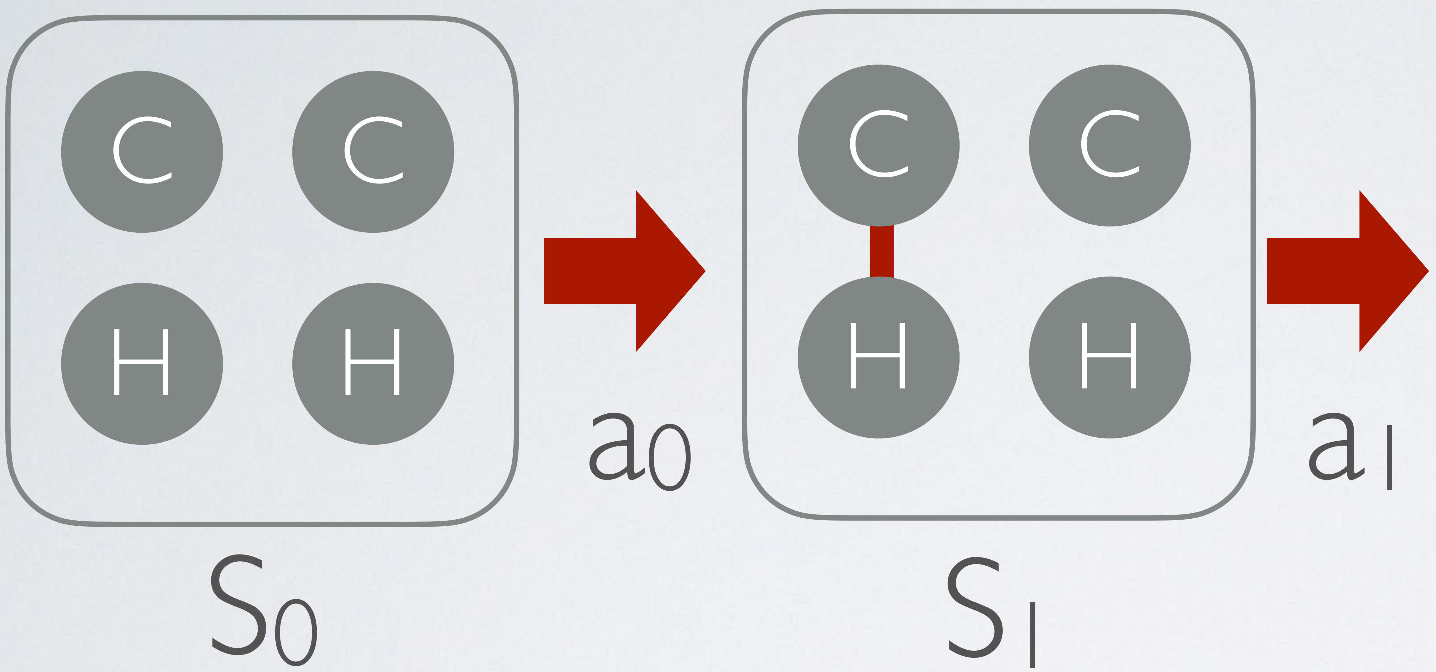
SEQUENTIAL BOND FORMATION



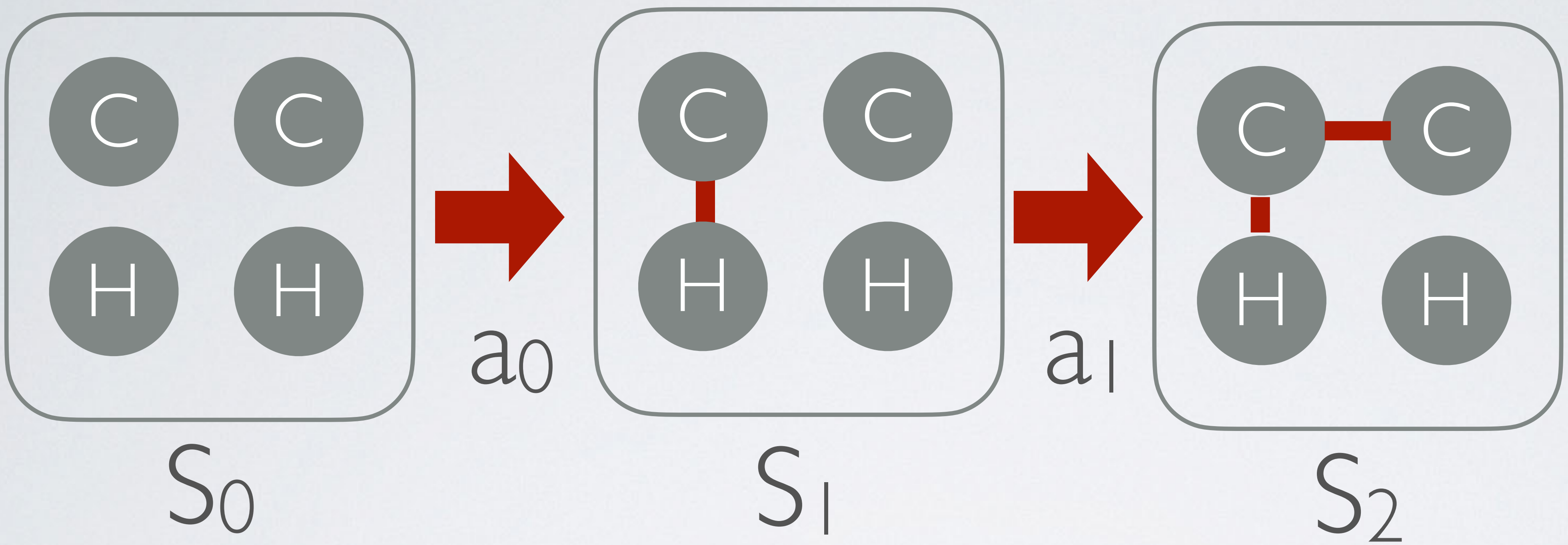
SEQUENTIAL BOND FORMATION



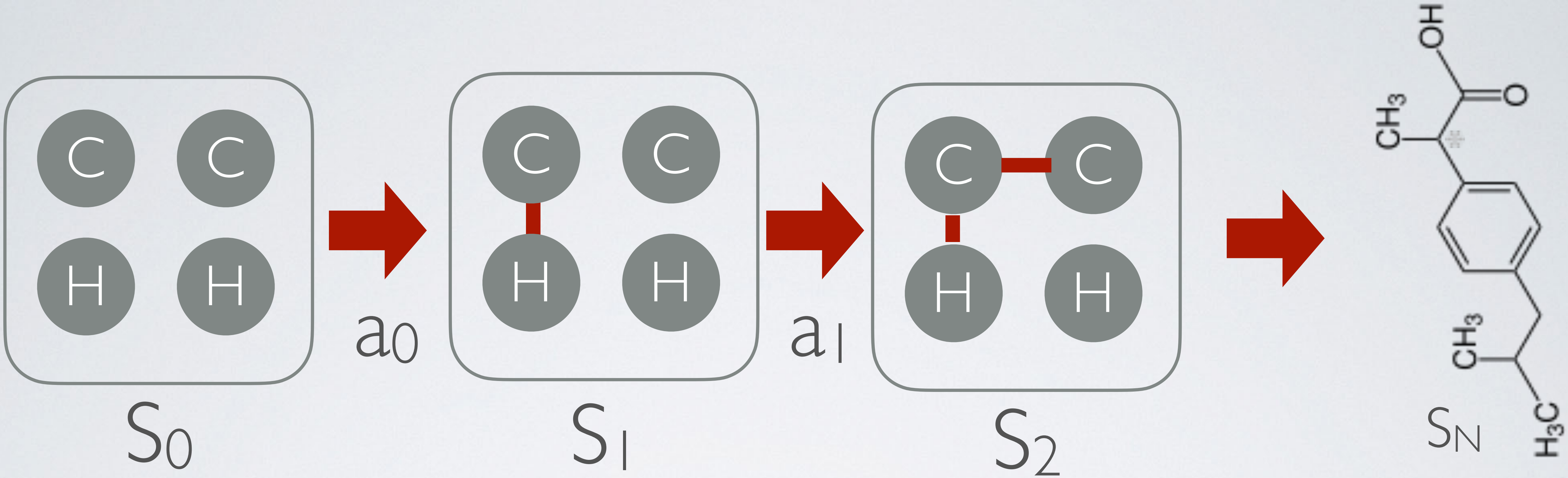
SEQUENTIAL BOND FORMATION



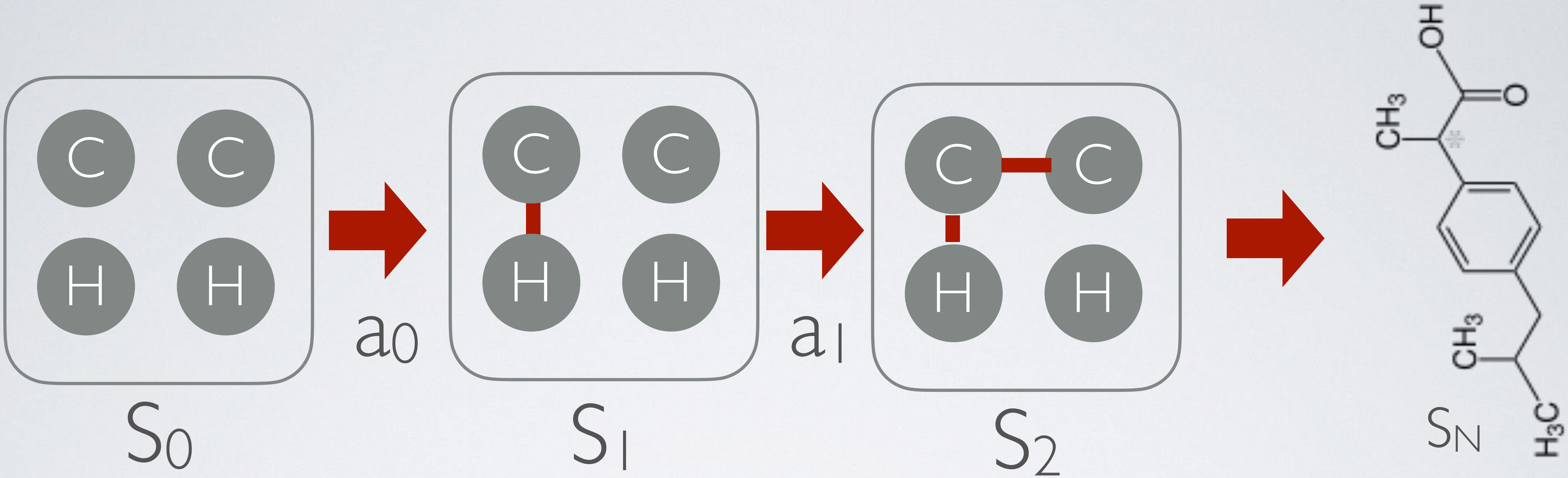
SEQUENTIAL BOND FORMATION



SEQUENTIAL BOND FORMATION

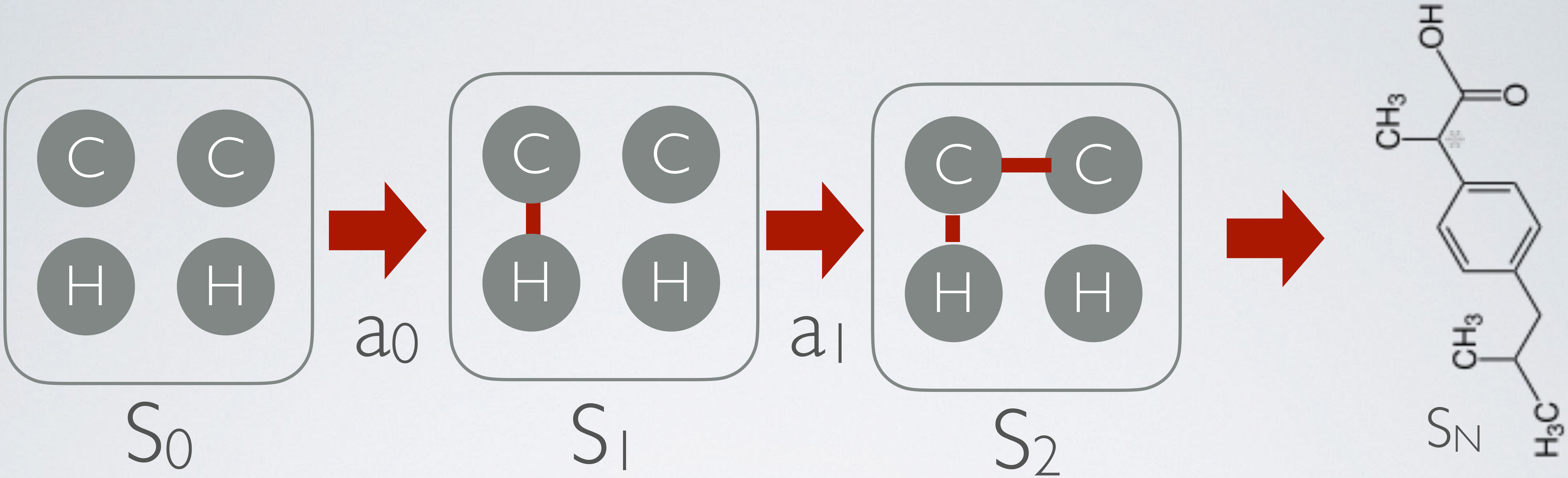


SEQUENTIAL BOND FORMATION



$p(a_k \mid S_k, \text{spectrum})$

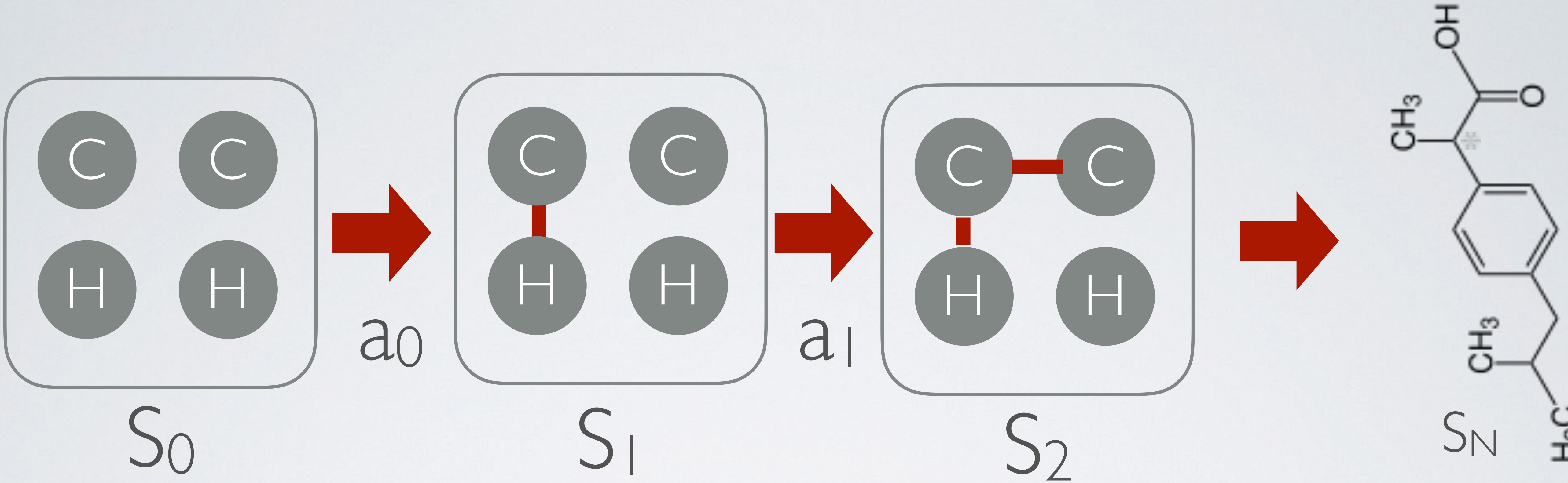
SEQUENTIAL BOND FORMATION



$p(a_k \mid S_k, \text{spectrum})$

Fine, this is a MDP.
Is it easy or hard?

SEQUENTIAL BOND FORMATION

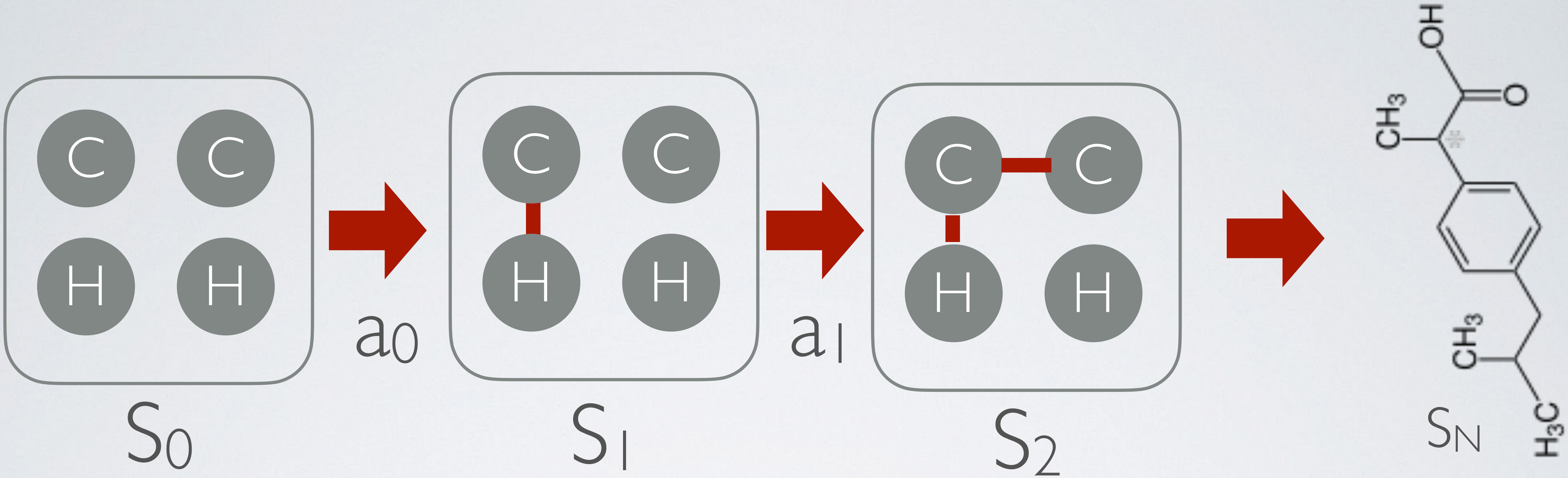


$p(a_k \mid S_k, \text{spectrum})$

Fine, this is a MDP.
Is it easy or hard?

- Easy: fully-observed

SEQUENTIAL BOND FORMATION

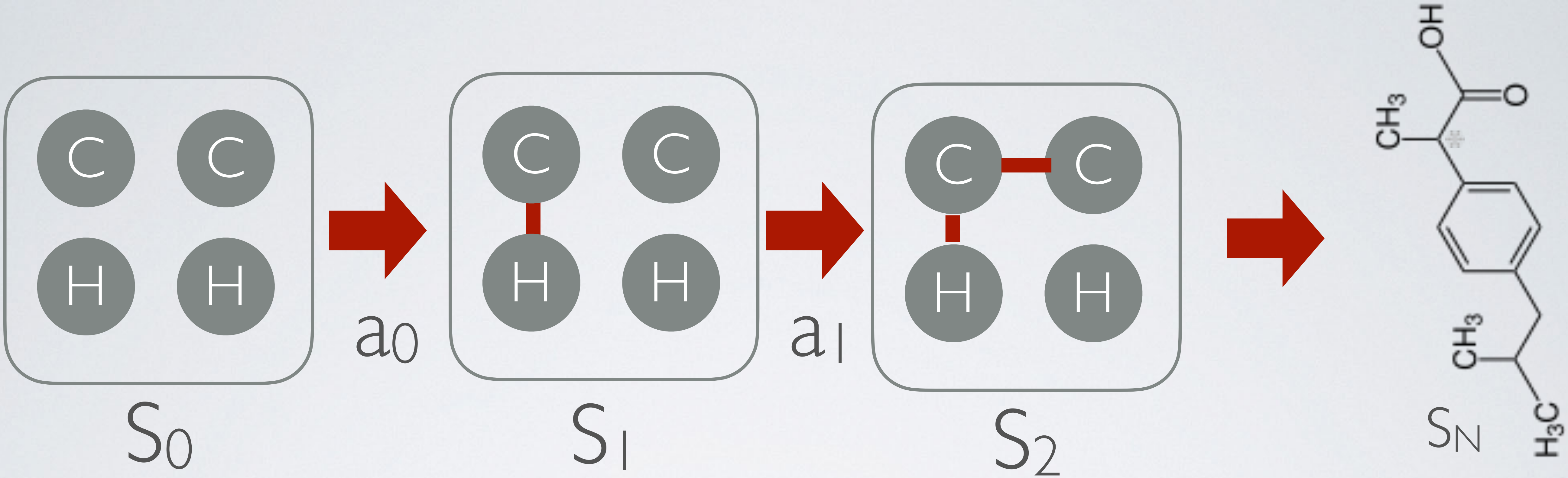


$$p(a_k \mid S_k, \text{spectrum})$$

Fine, this is a MDP.
Is it easy or hard?

- Easy: fully-observed
- Easy: Deterministic actions

SEQUENTIAL BOND FORMATION

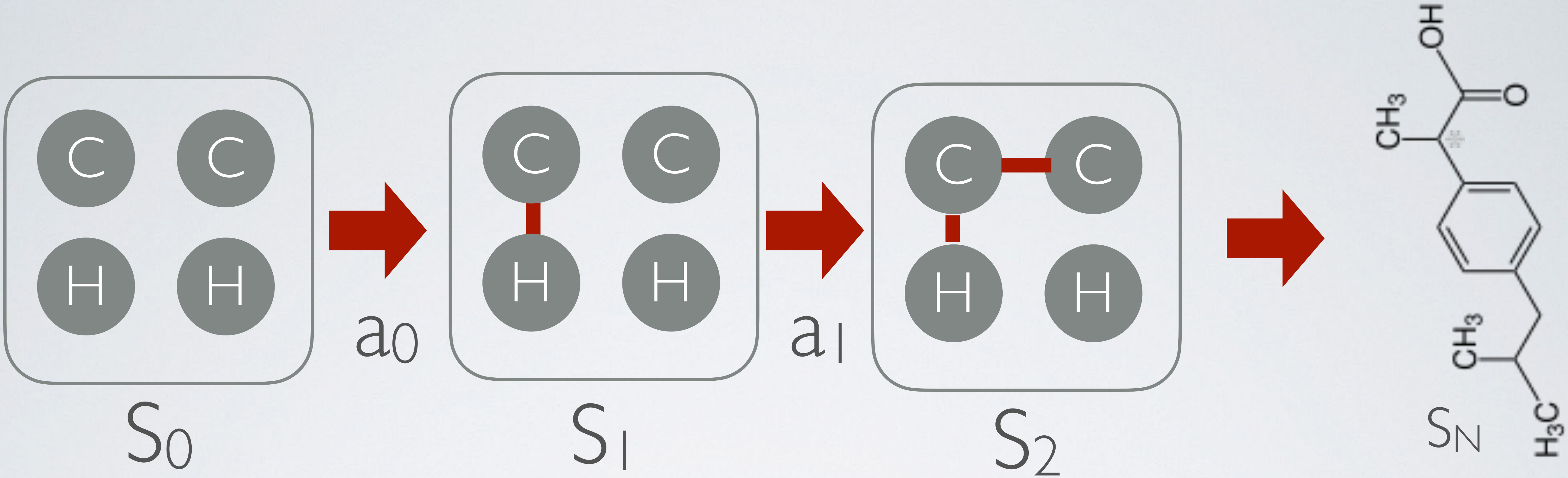


$p(a_k \mid S_k, \text{spectrum})$

Fine, this is a MDP.
Is it easy or hard?

- Easy: fully-observed
- Easy: Deterministic actions
- Hard: delayed reward

SEQUENTIAL BOND FORMATION



$p(a_k \mid S_k, \text{spectrum})$

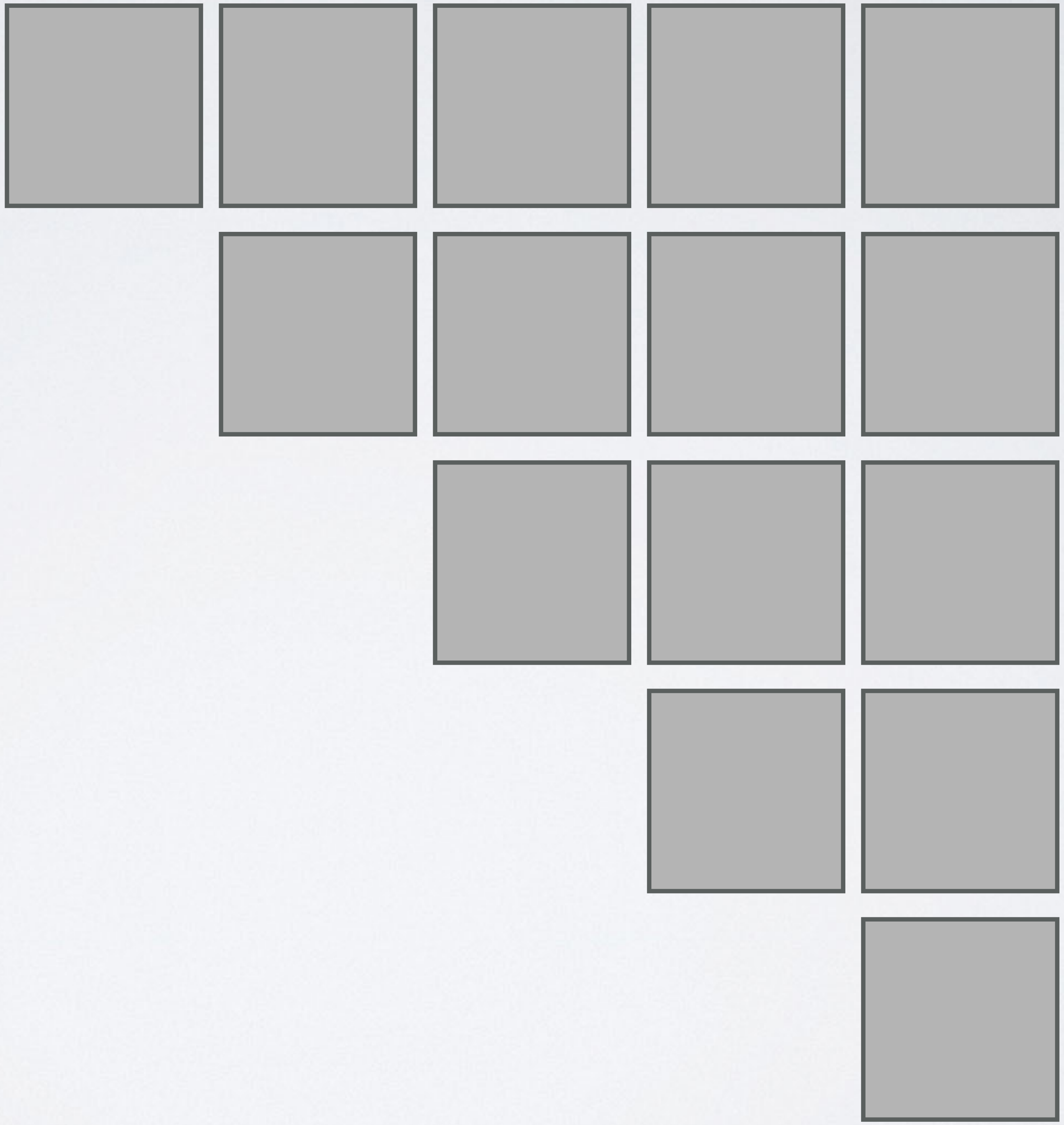
Fine, this is a MDP.
Is it easy or hard?

- Easy: fully-observed
- Easy: Deterministic actions
- Hard: delayed reward

Treat as an imitation learning problem

WHAT IS THE RIGHT NEXT ACTION?

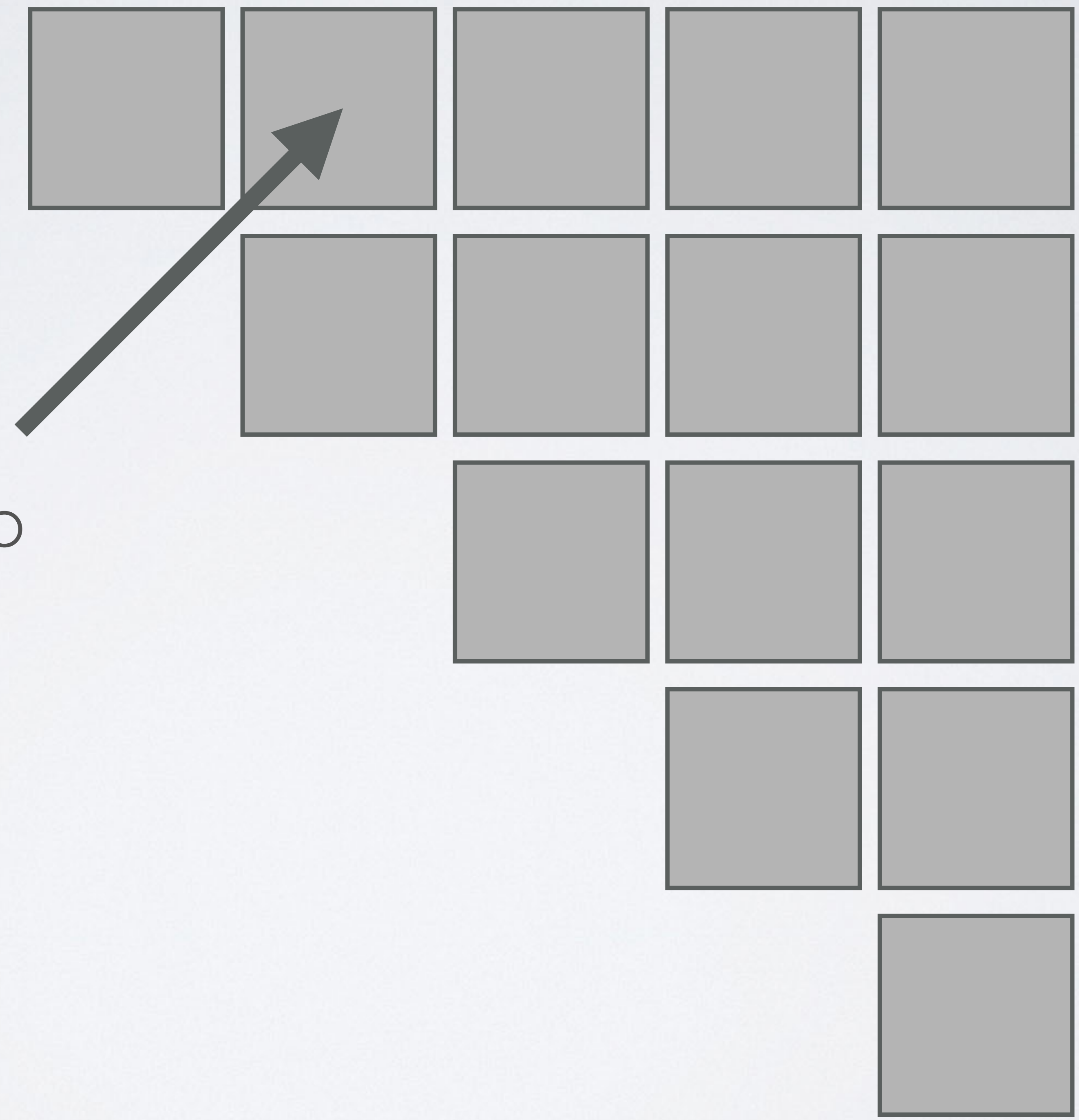
At training time, we have an **oracle** which tells us all valid next edges



WHAT IS THE RIGHT NEXT ACTION?

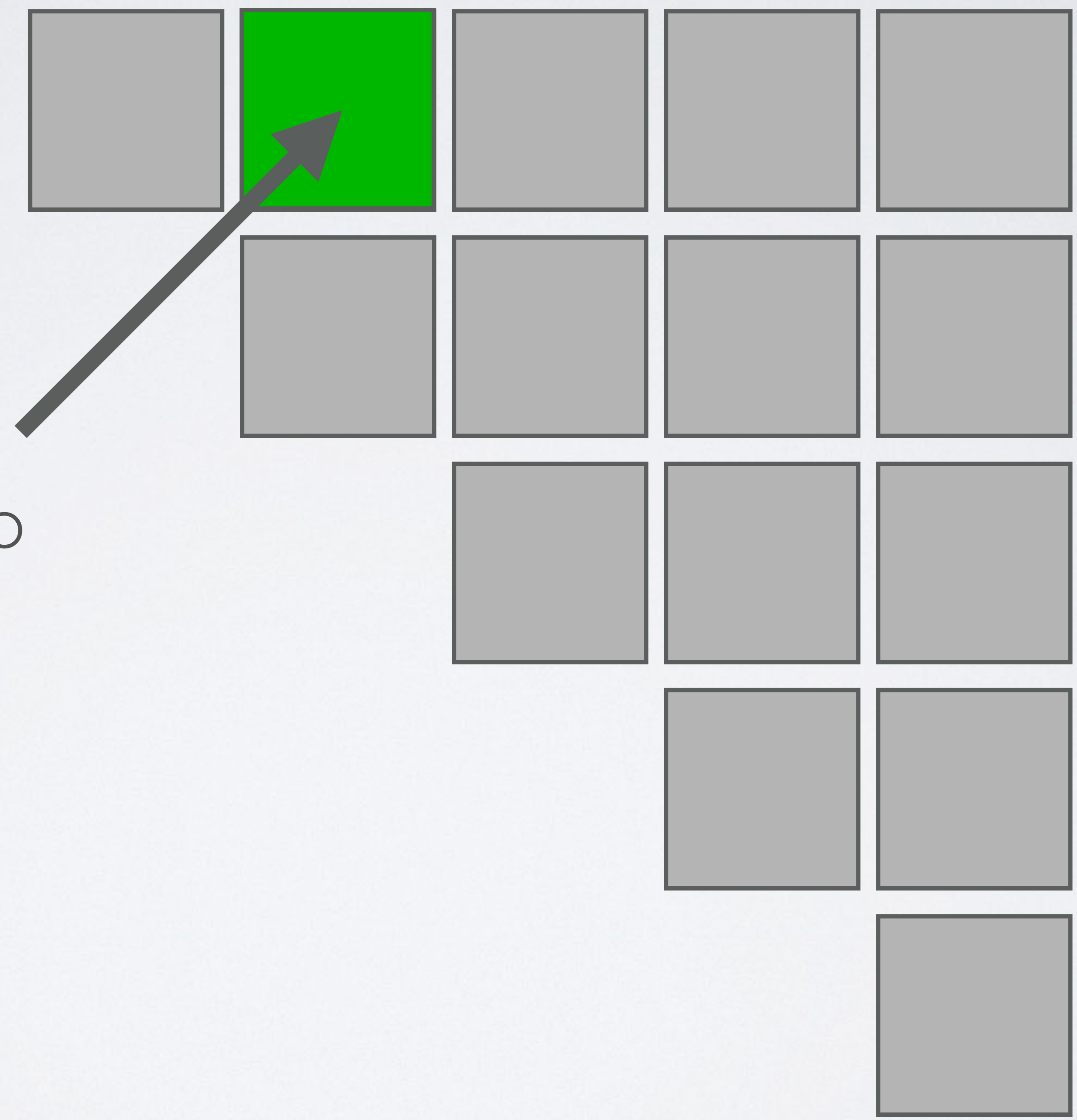
At training time, we have an **oracle** which tells us all valid next edges

If we add this edge, is the resulting graph subisomorphic to the truth? **YES**



WHAT IS THE RIGHT NEXT ACTION?

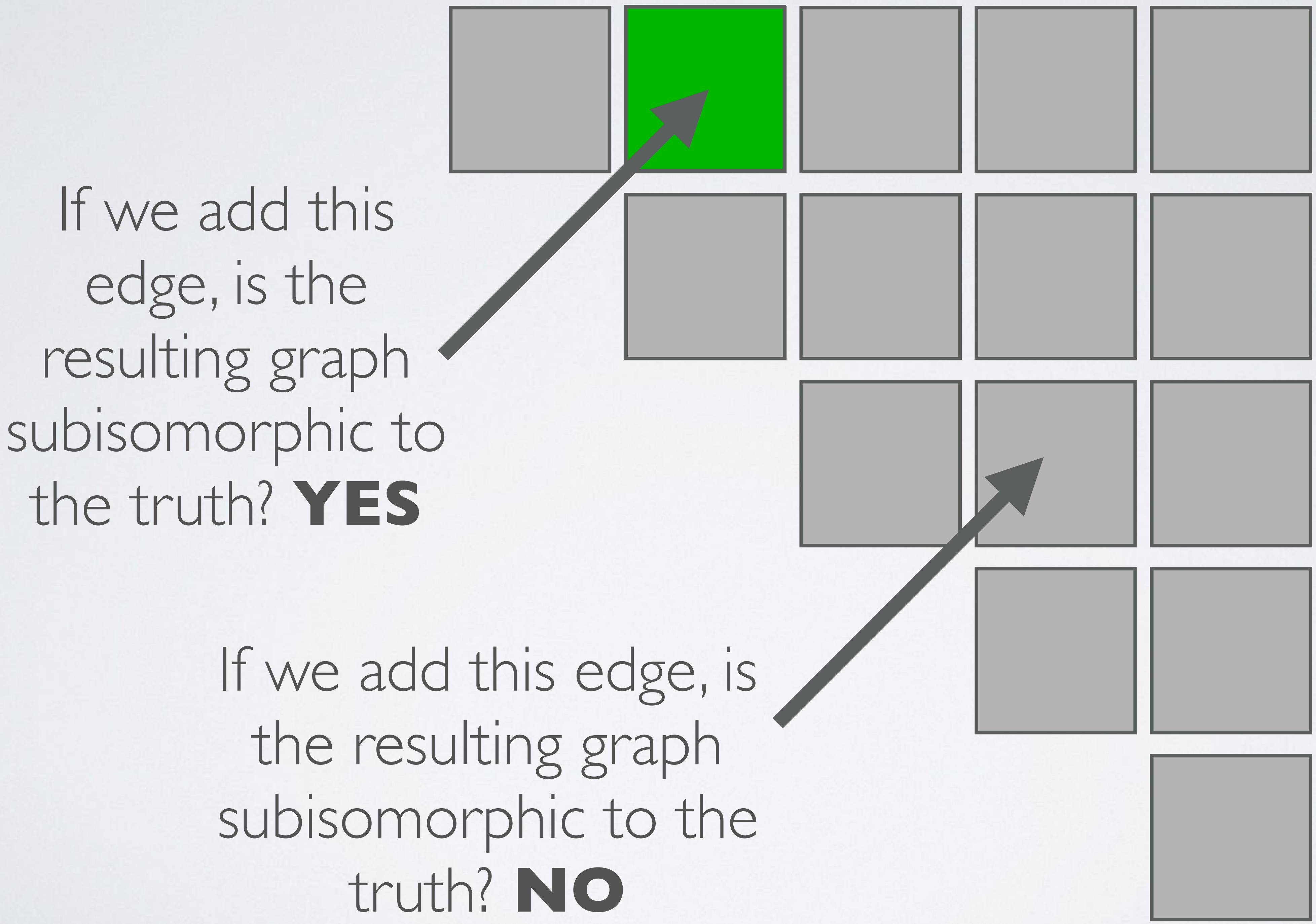
At training time, we have an **oracle** which tells us all valid next edges



If we add this edge, is the resulting graph subisomorphic to the truth? **YES**

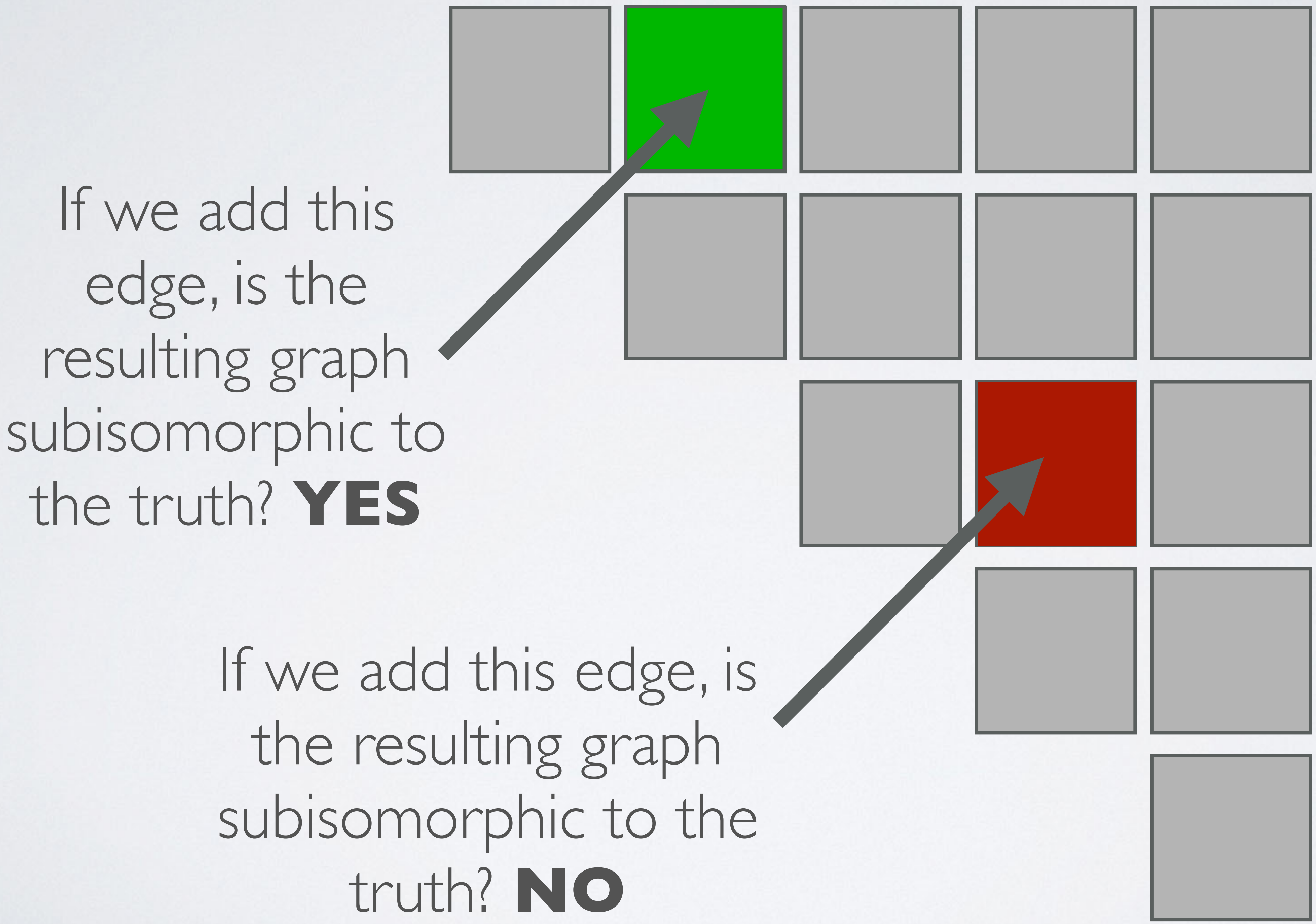
WHAT IS THE RIGHT NEXT ACTION?

At training time, we have an **oracle** which tells us all valid next edges



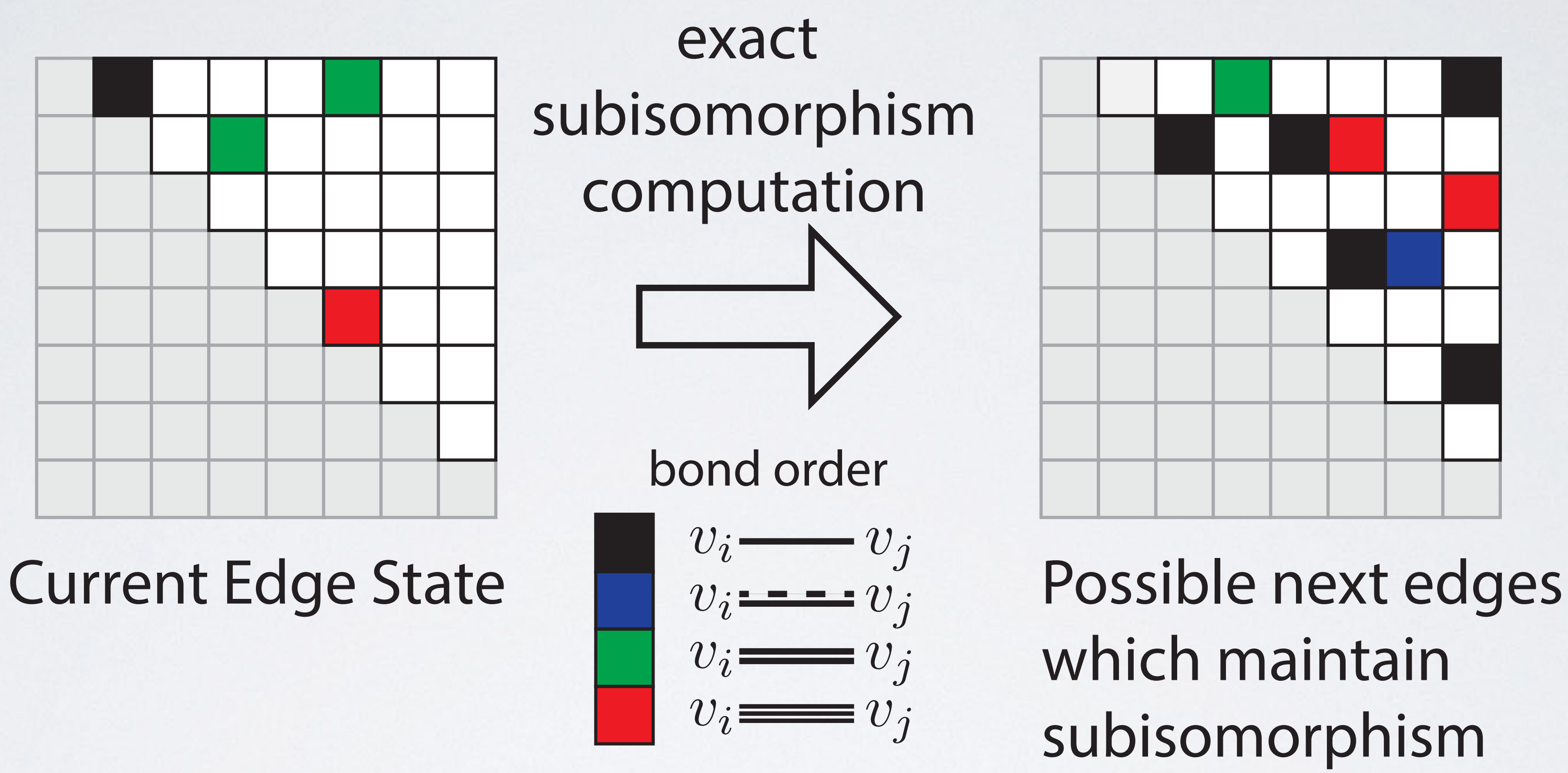
WHAT IS THE RIGHT NEXT ACTION?

At training time, we have an **oracle** which tells us all valid next edges



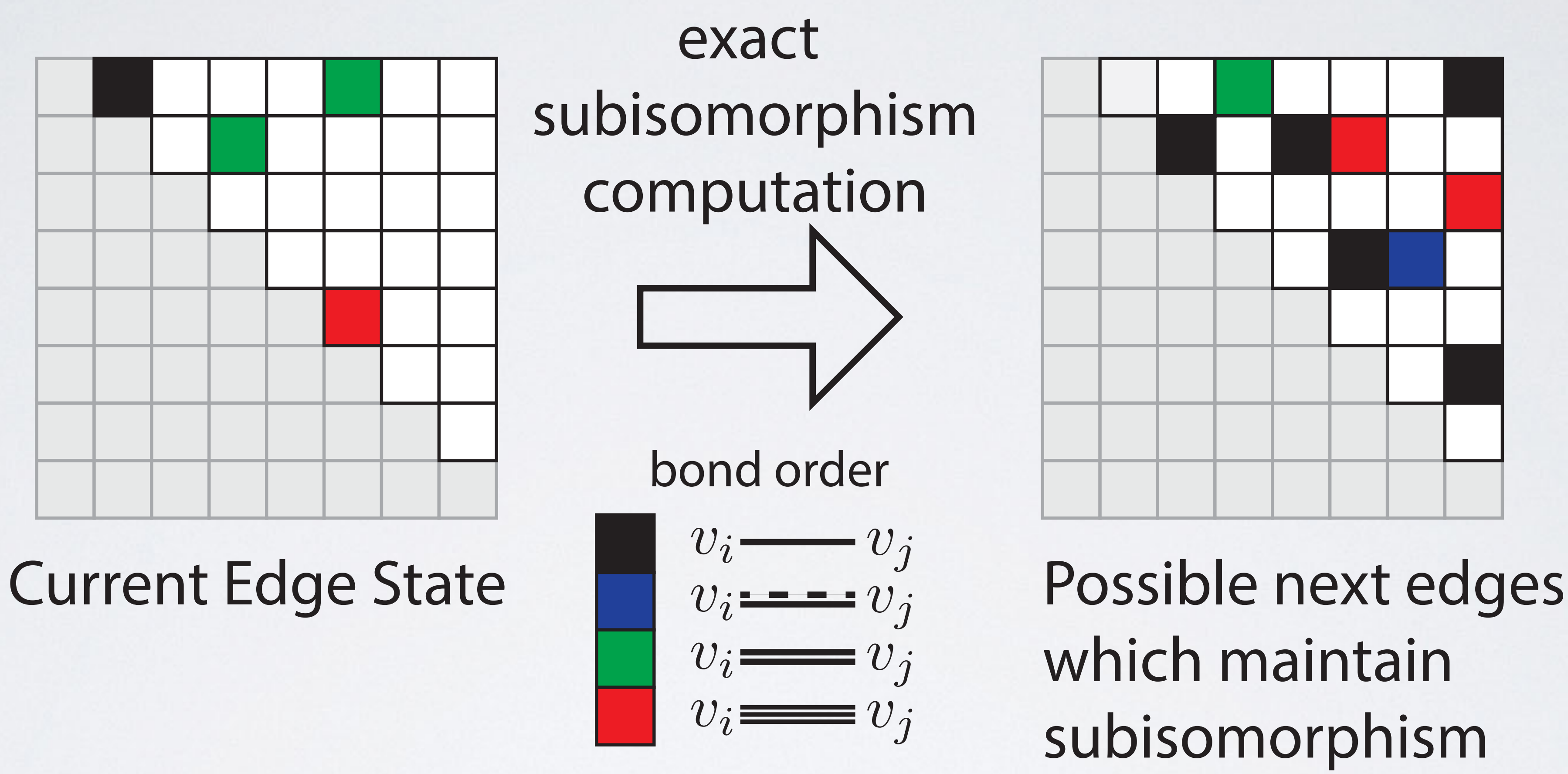
WHAT IS THE RIGHT NEXT ACTION?

At training time, we have an **oracle** which tells us all valid next edges



WHAT IS THE RIGHT NEXT ACTION?

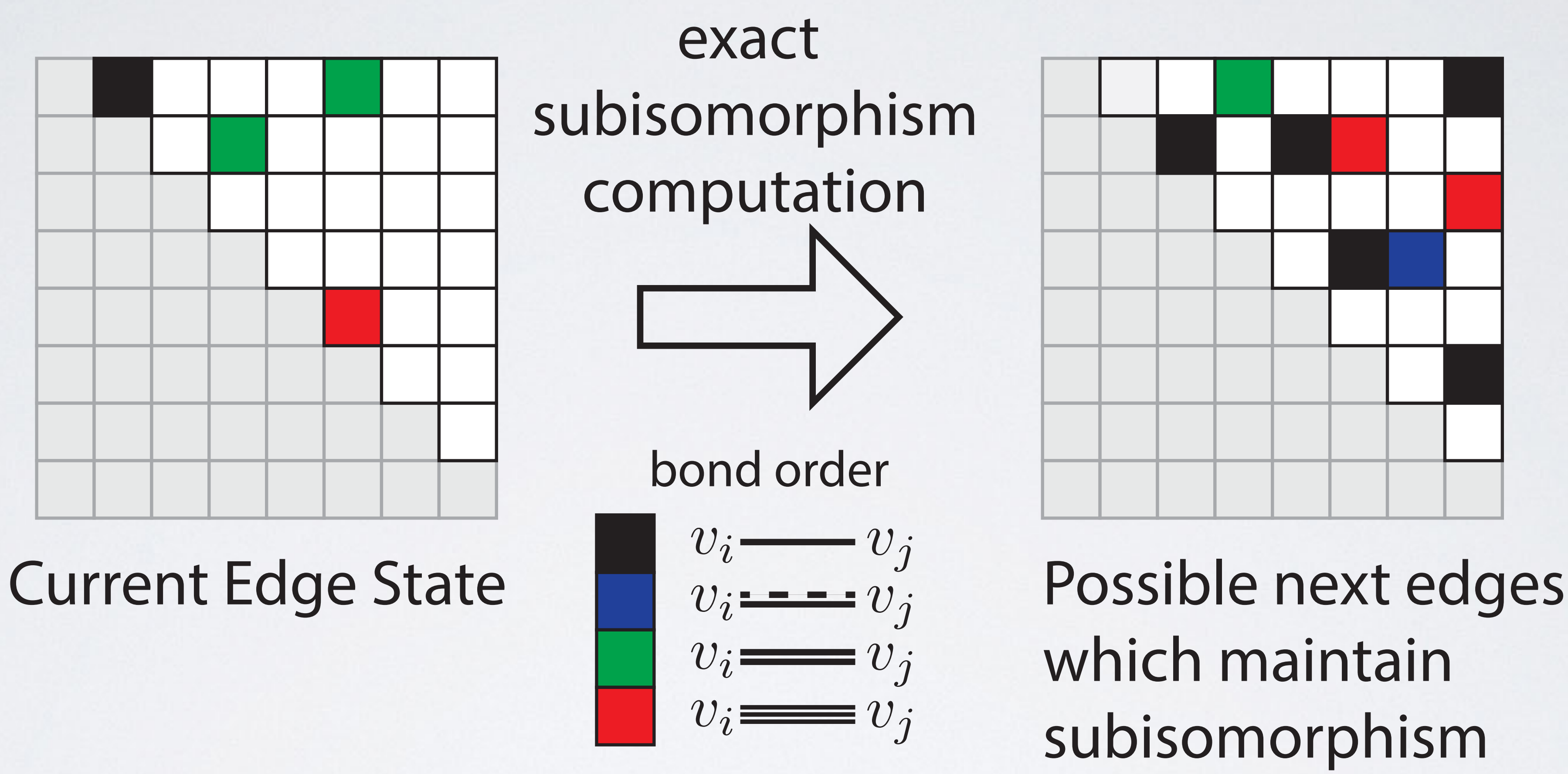
At training time, we have an **oracle** which tells us all valid next edges



$$\min_{\theta} ||A_{\mathrm{subiso}} - p_{\theta}(A|S_t, spectrum)||$$

WHAT IS THE RIGHT NEXT ACTION?

At training time, we have an **oracle** which tells us all valid next edges



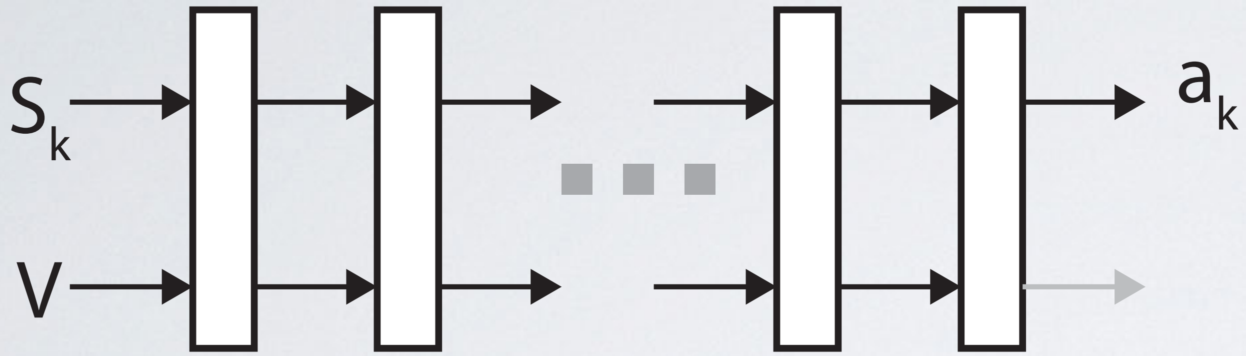
$$\min_{\theta} ||A_{\mathrm{subiso}} - p_{\theta}(A|S_t, spectrum)||$$

We can thus directly minimize policy over valid actions — imitation learning

$p(a_k \mid S_k, \text{spectrum})$

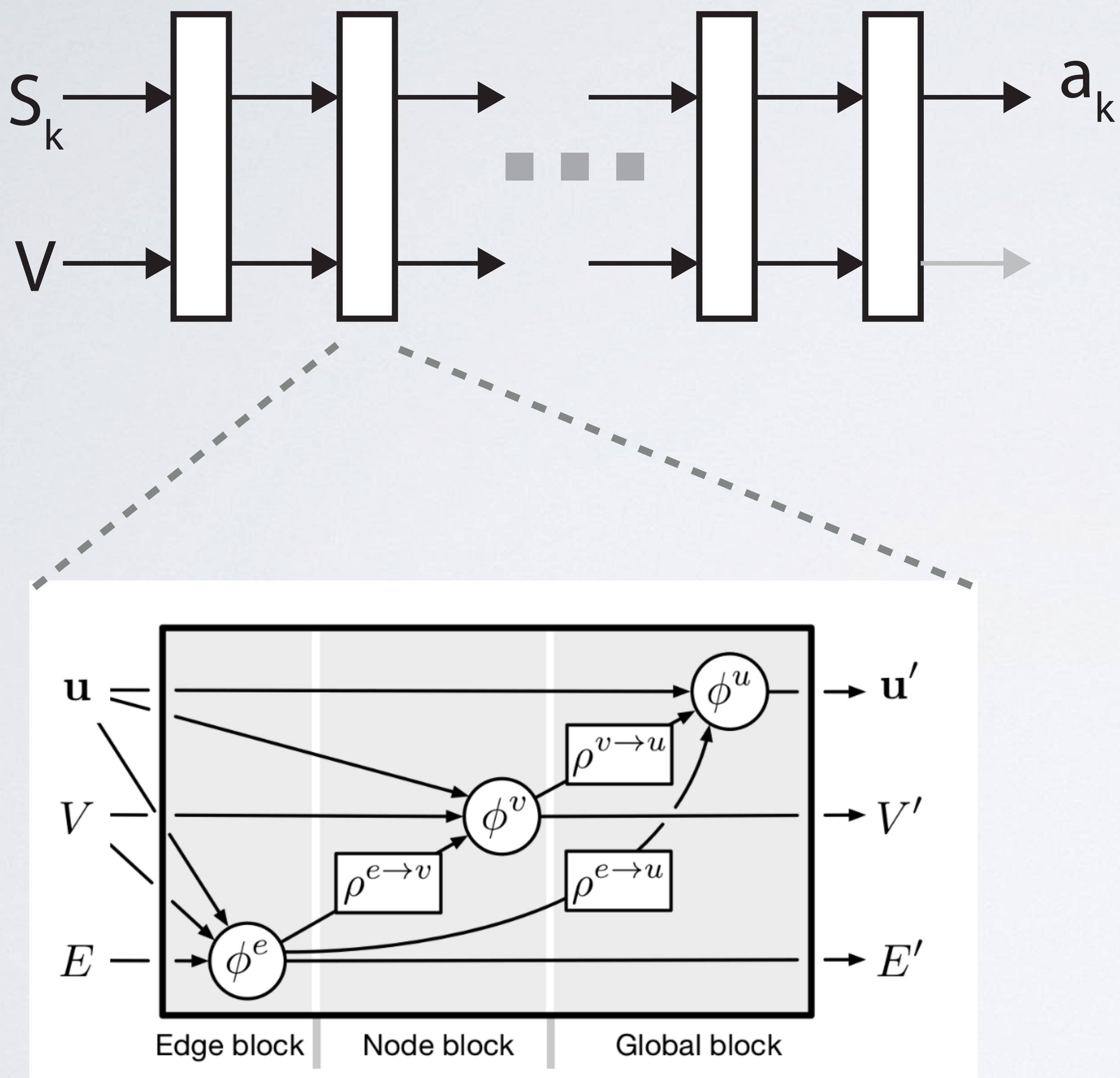
$p(a_k \mid S_k, \text{spectrum})$

Relational network



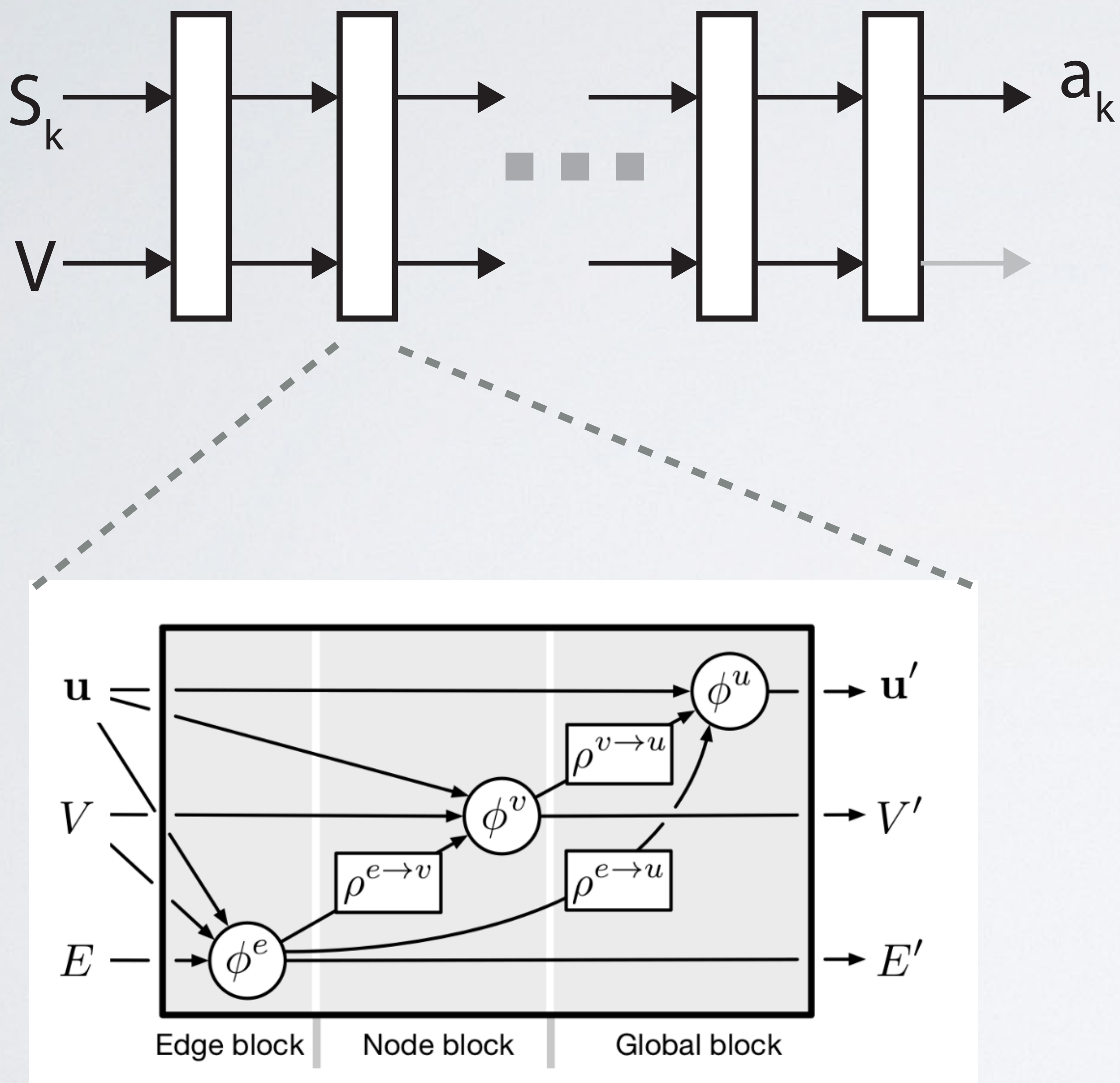
$p(a_k \mid S_k, \text{spectrum})$

Relational network



$p(a_k \mid S_k, \text{spectrum})$

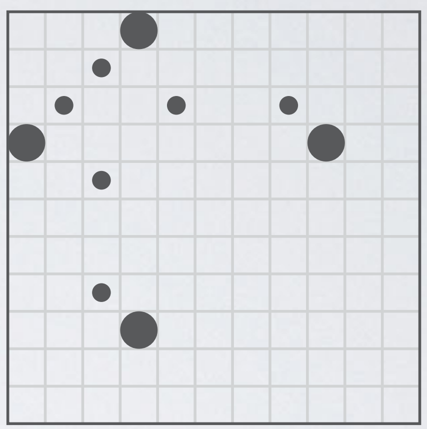
Relational network



Training data

	freq	valence
C	72	4
C	43	4
C	107	4
C	144	4
O	-	2
O	-	2
N	-	3
H	-	1
H	-	1
H	-	1
H	-	1

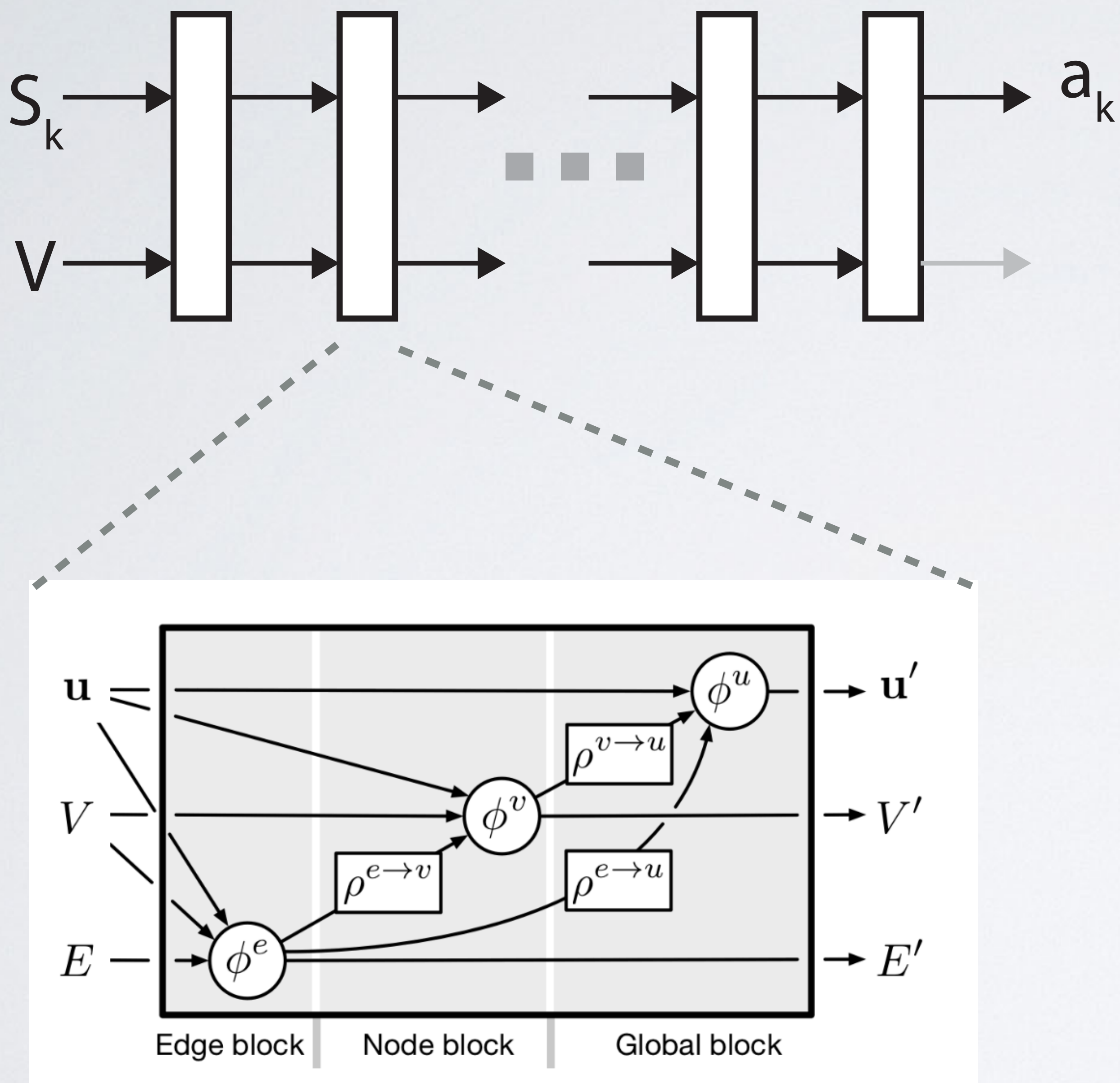
per-vertex data



S_k

$p(a_k \mid S_k, \text{spectrum})$

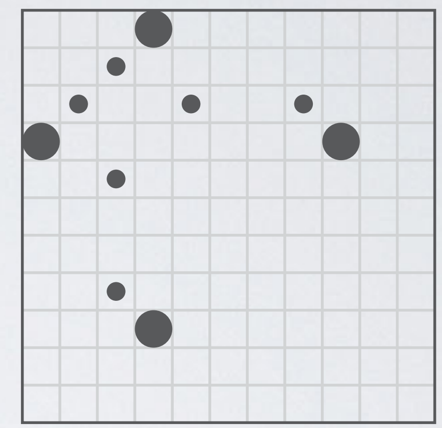
Relational network



Training data

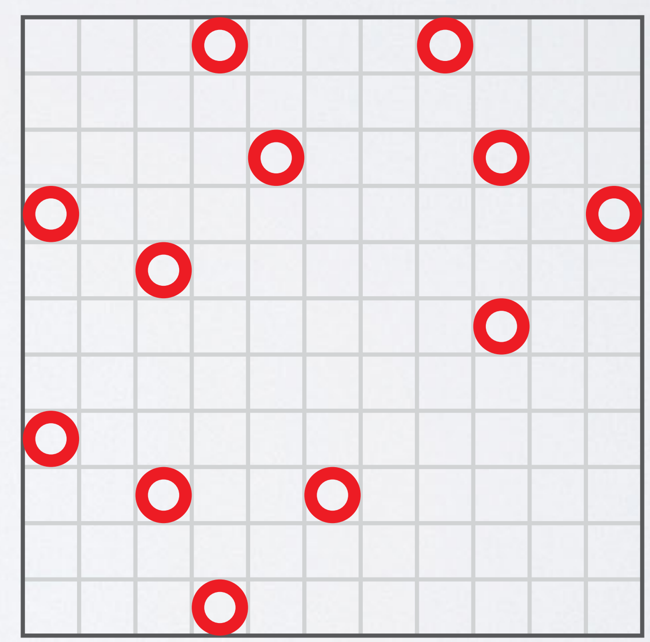
	freq	valence
C	72	4
C	43	4
C	107	4
C	144	4
O	-	2
O	-	2
N	-	3
H	-	1
H	-	1
H	-	1
H	-	1

per-vertex data



S_k

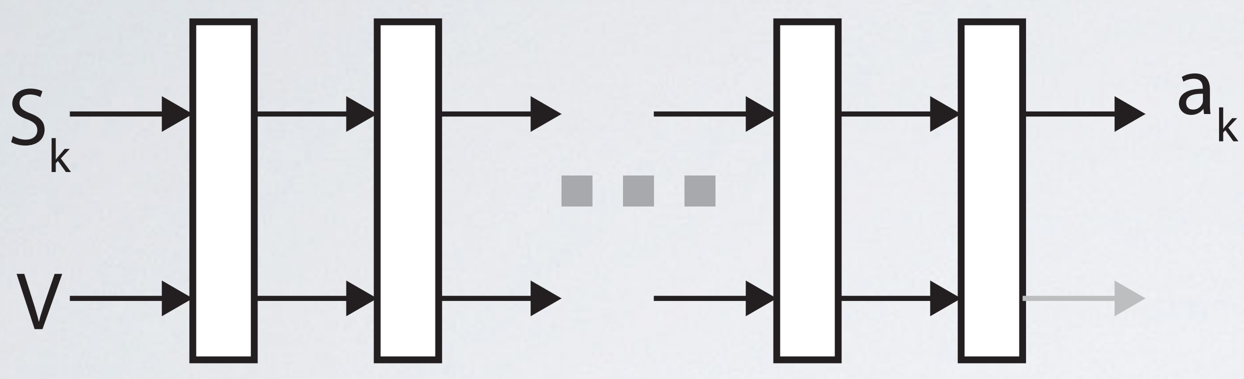
All valid next edges

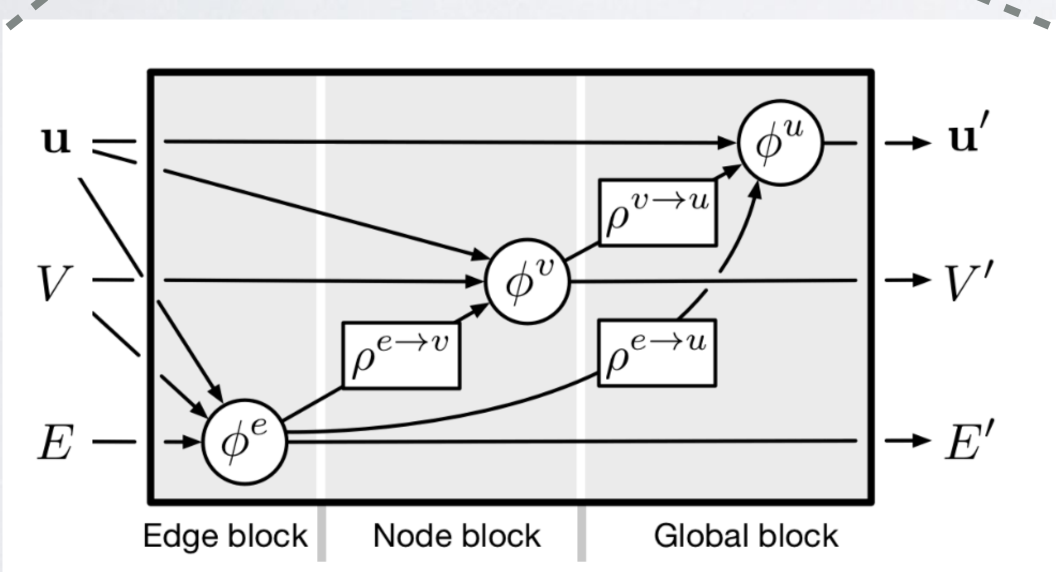


a_k

p(a_k | S_k, spectrum)

Relational network



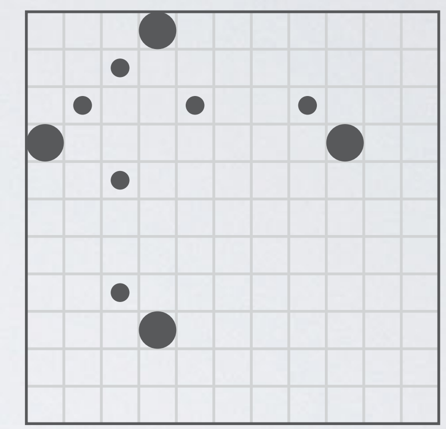


Train on 1.3M molecules in PubChem
HCON, up to 32 heavy atoms
13C chemical shifts

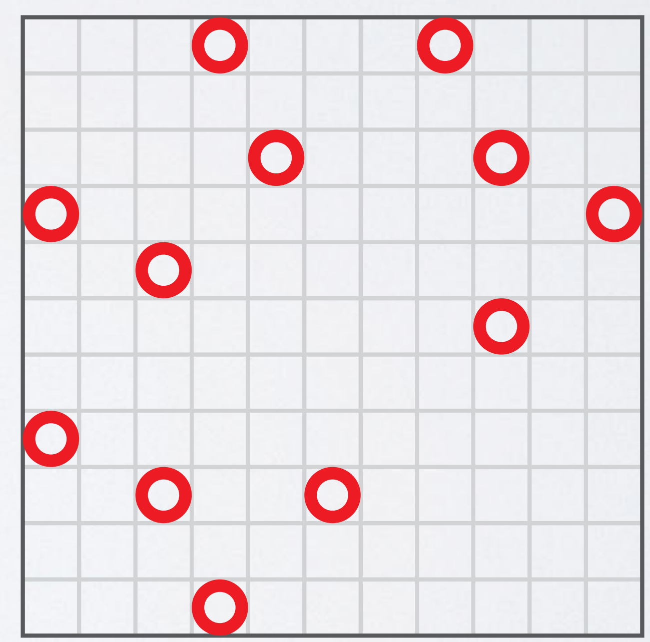
Training data

	freq	valence
C	72	4
C	43	4
C	107	4
C	144	4
O	-	2
O	-	2
N	-	3
H	-	1
H	-	1
H	-	1
H	-	1

per-vertex data



S_k



All valid next edges

a_k

Say I've learned:

$p(a_k \mid S_k, \text{spectrum})$

What happens at test time?

Say I've learned:

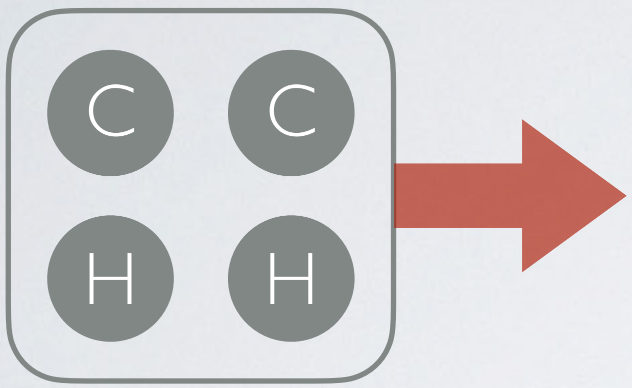
$p(a_k \mid S_k, \text{spectrum})$

What happens at test time?

132.4 73.1

Say I've learned:

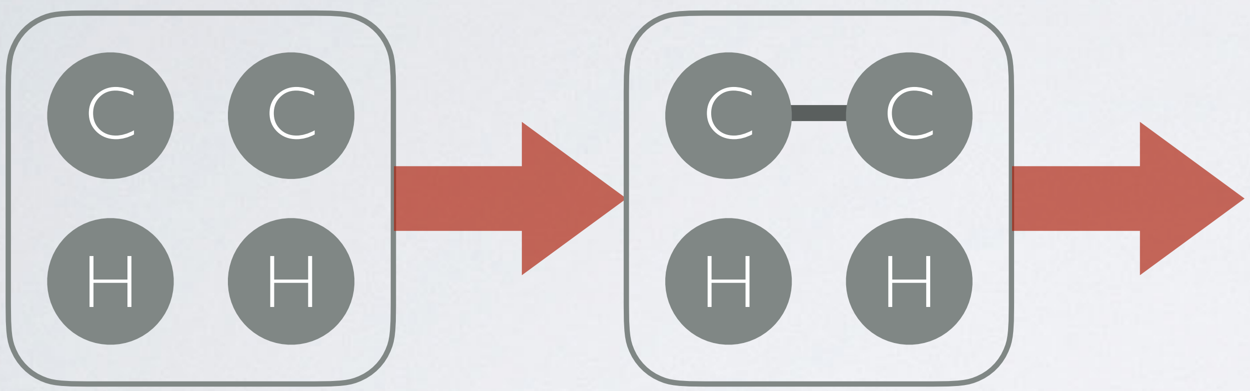
$p(a_k \mid S_k, \text{spectrum})$



What happens at test time?

Say I've learned:

$p(a_k \mid S_k, \text{spectrum})$

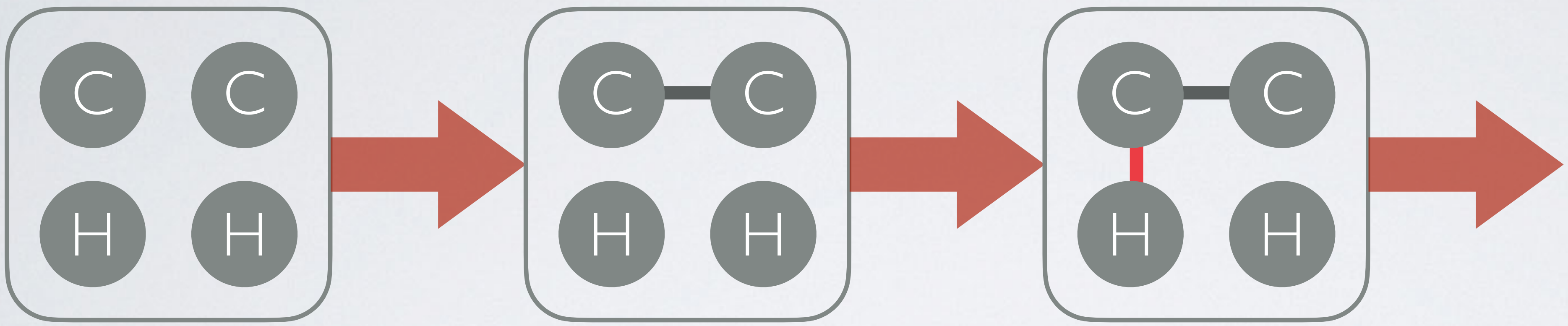


What happens at test time?

Say I've learned:

$$p(a_k \mid S_k, \text{spectrum})$$

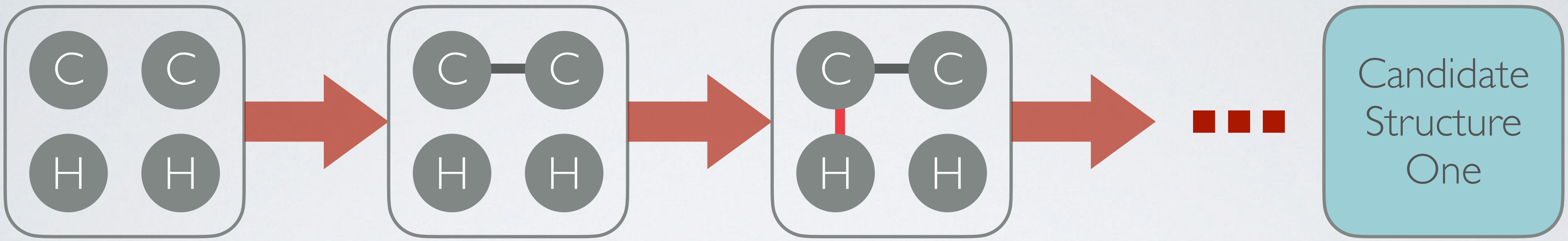
What happens at test time?



Say I've learned:

$p(a_k \mid S_k, \text{spectrum})$

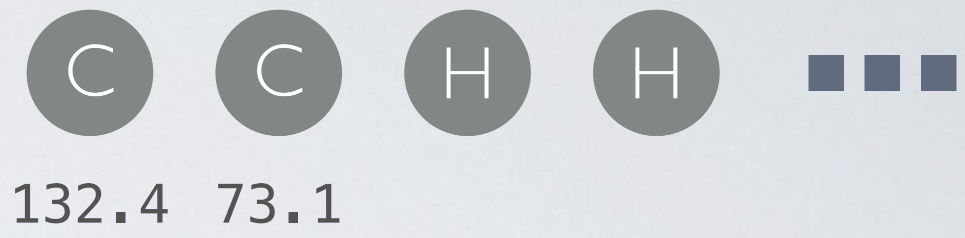
What happens at test time?

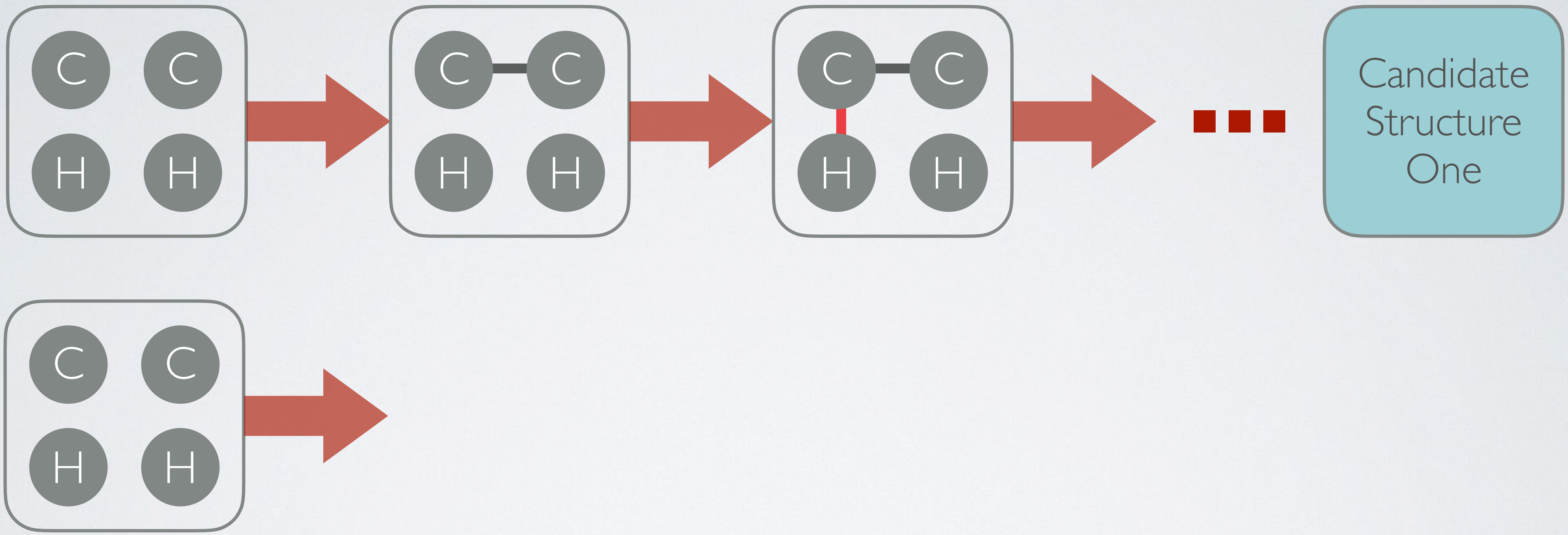


Say I've learned:

$$p(a_k \mid S_k, \text{spectrum})$$

What happens at test time?

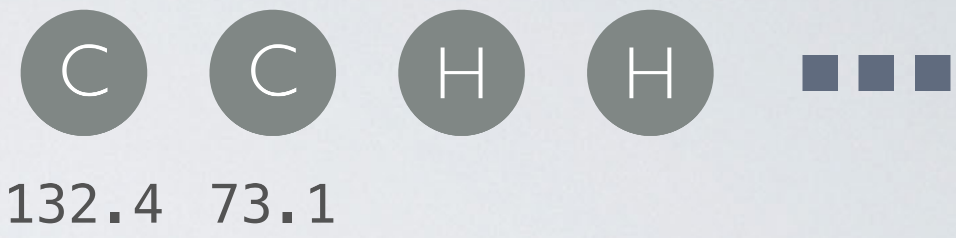


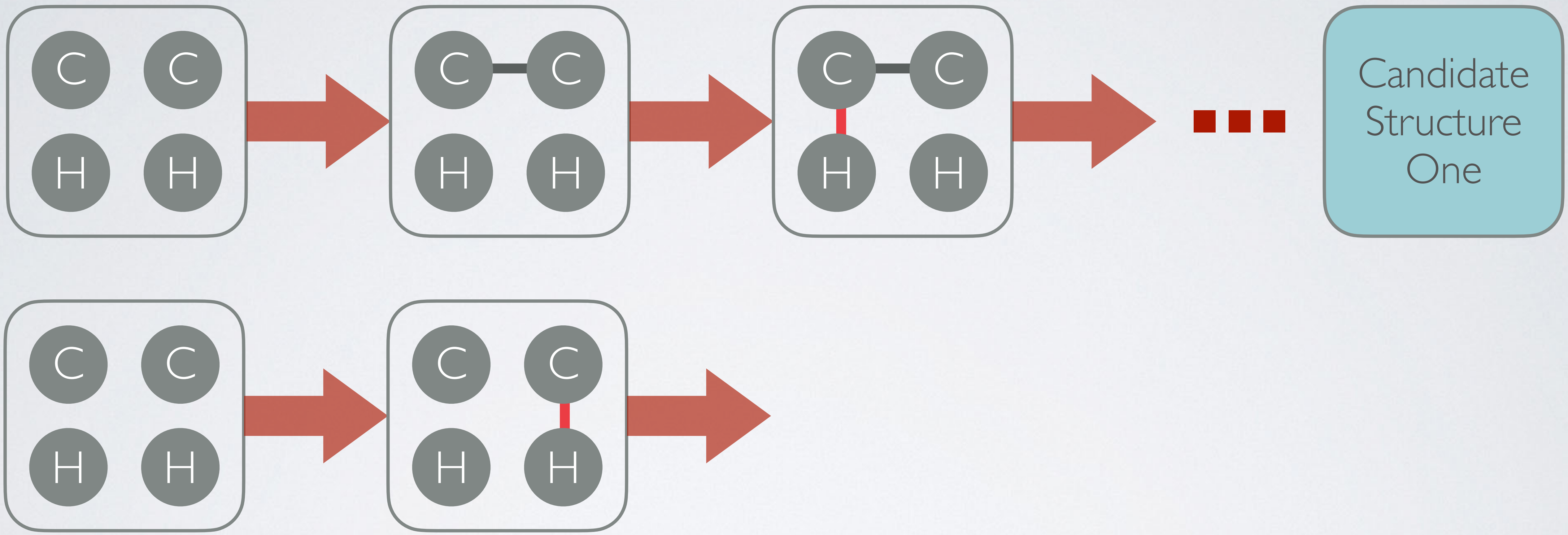


Say I've learned:

$$p(a_k \mid S_k, \text{spectrum})$$

What happens at test time?

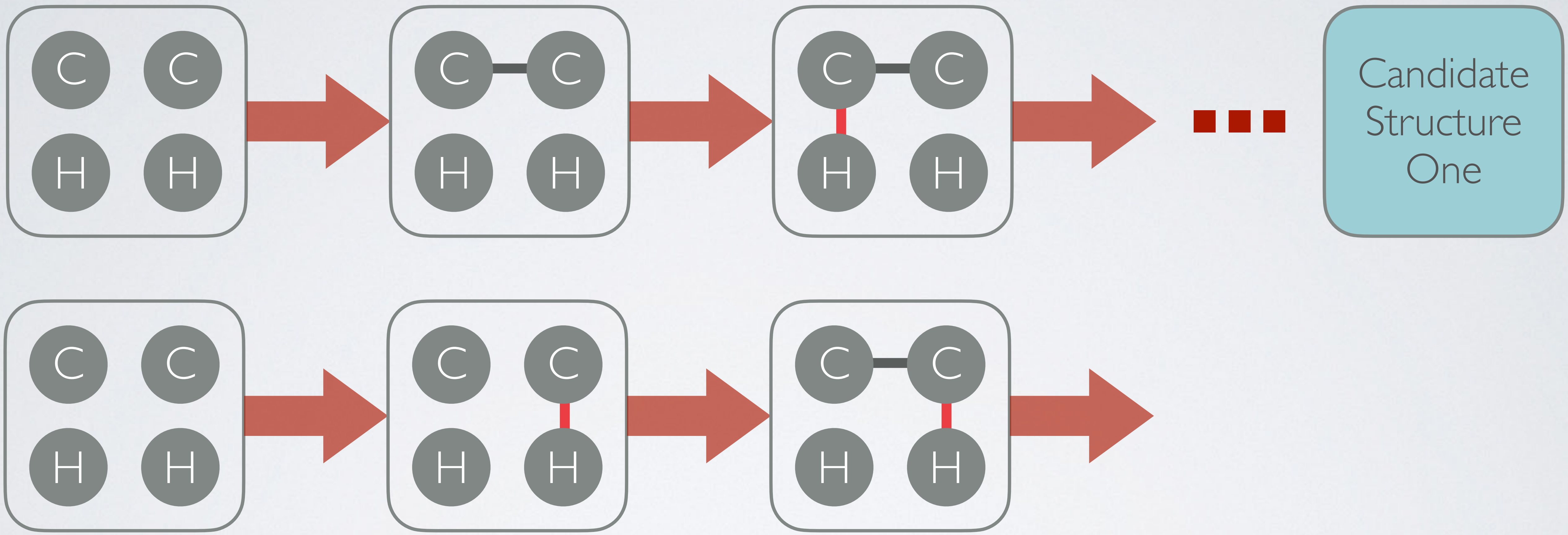




Say I've learned:

$$p(a_k \mid S_k, \text{spectrum})$$

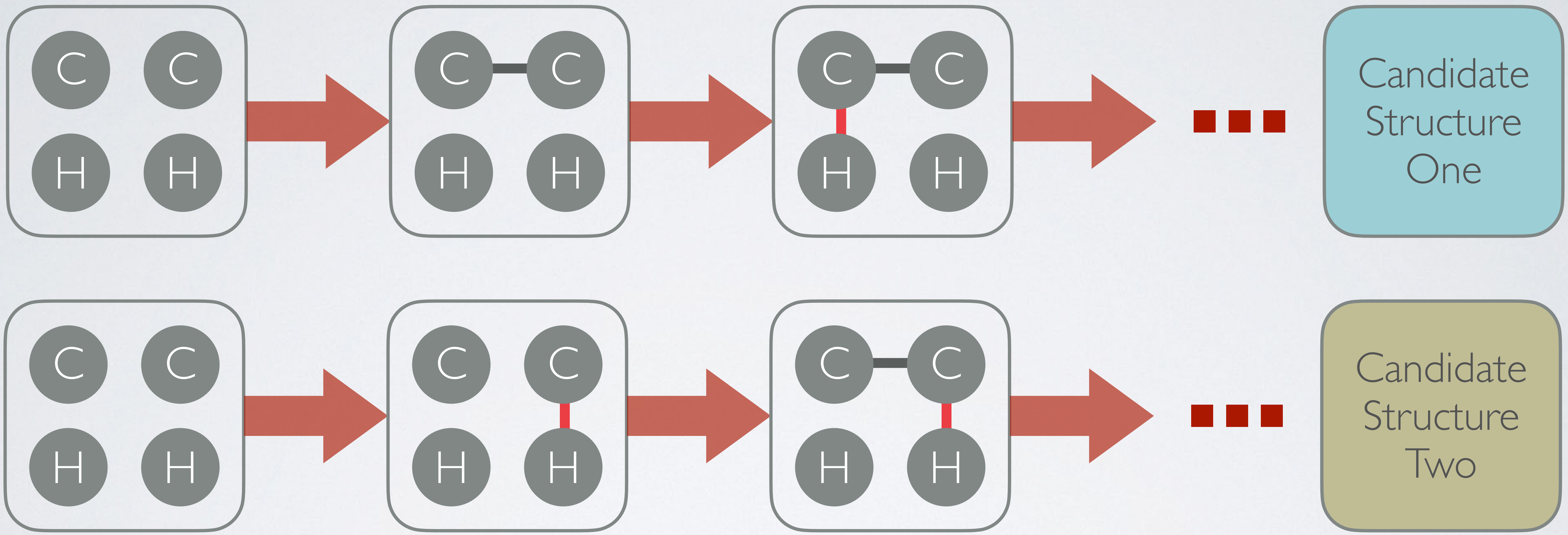
What happens at test time?



Say I've learned:

$$p(a_k \mid S_k, \text{spectrum})$$

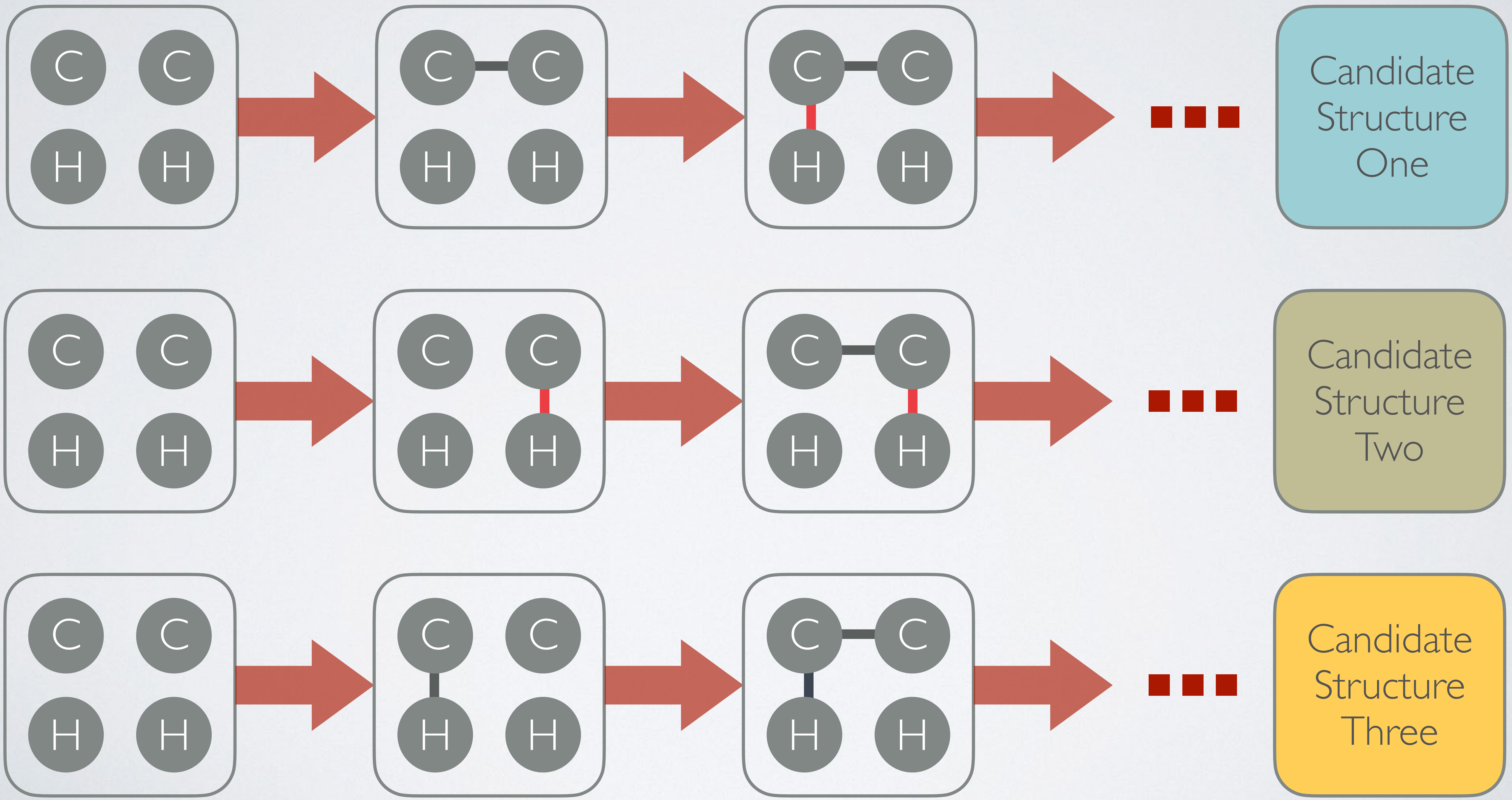
What happens at test time?



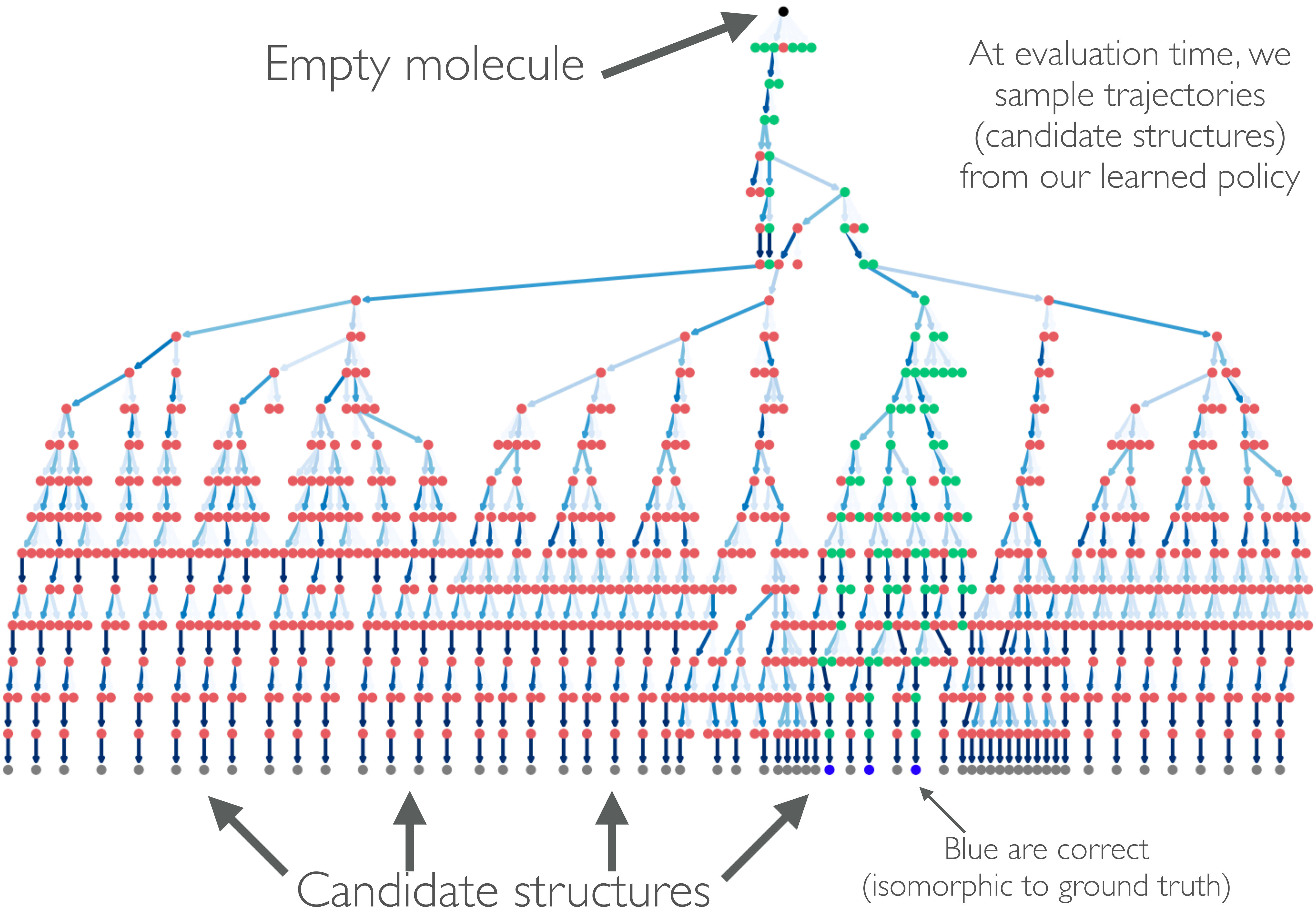
Say I’ve learned:

$$p(a_k \mid S_k, \text{spectrum})$$

What happens at test time?



GENERATING MOLECULES



At evaluation time, we sample trajectories (candidate structures) from our learned policy

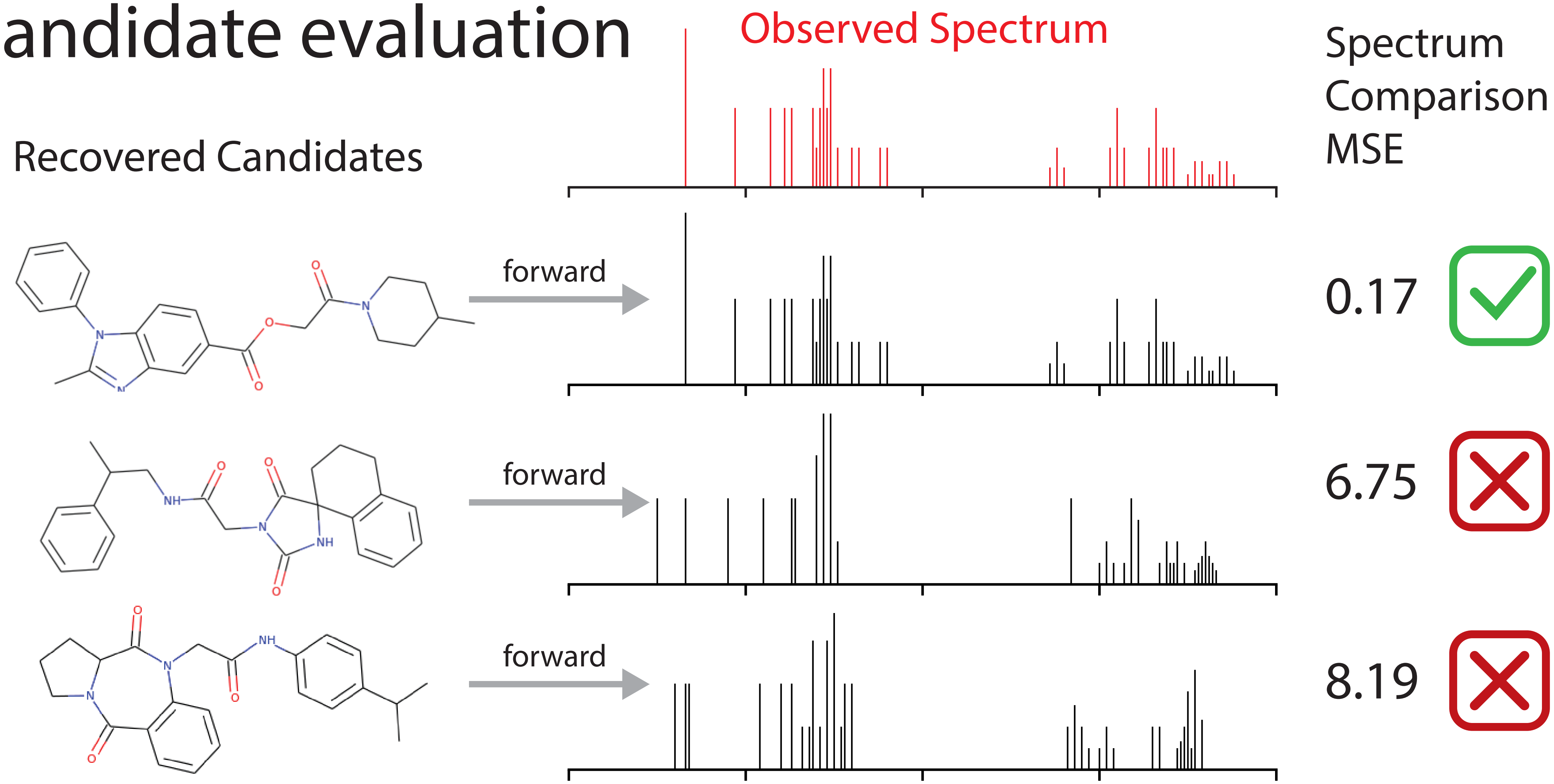
PREDICTION WITH CONFIDENCE

Is there any way to know which of our candidate structures is correct?

PREDICTION WITH CONFIDENCE

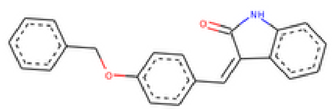
Is there any way to know which of our candidate structures is correct?

Candidate evaluation

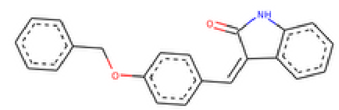


Example Reconstructions

true structure | **candidate structures**

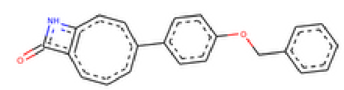


O=C1Nc2ccccc2C1=Cc1ccc(OCc2ccccc2)cc1



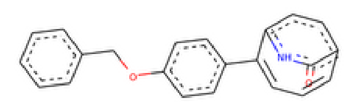
0.0

O=C1Nc2ccccc2C1=Cc1ccc(OCc2ccccc2)cc1



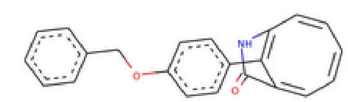
57.2

O=c1[nH]c2ccc(-c3ccc(OCc4ccccc4)cc3)cccc1=2



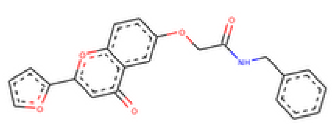
62.6

O=c1[nH]c2ccc1C=CC=C2c1ccc(OCc2ccccc2)cc1

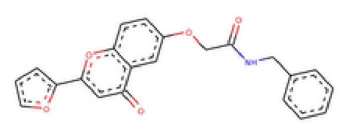


66.6

O=C1NC2=C(c3ccc(OCc4ccccc4)cc3)C1=CC=CC=C2

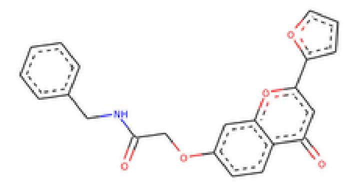


O=C(COc1ccc2oc(-c3ccco3)cc(=O)c2c1)NCc1ccccc1



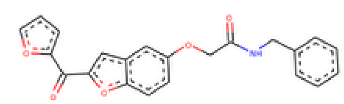
0.0

O=C(COc1ccc2oc(-c3ccco3)cc(=O)c2c1)NCc1ccccc1



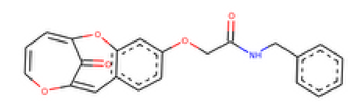
36.0

O=C(COc1ccc2c(=O)cc(-c3ccco3)oc2c1)NCc1ccccc1



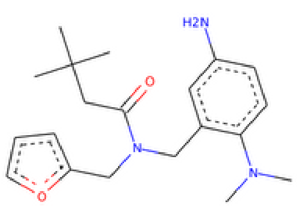
49.0

O=C(COc1ccc2oc(C(=O)c3ccco3)cc2c1)NCc1ccccc1

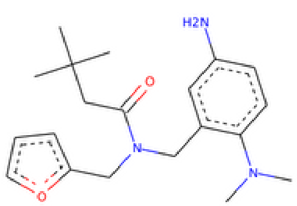


53.4

O=C(COc1ccc2c(c1)OC1=CC=COC(=C2)C1=O)NCc1ccccc1

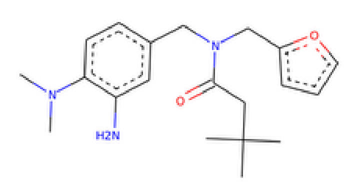


CN(C)c1ccc(N)cc1CN(Cc1ccco1)C(=O)CC(C)(C)C



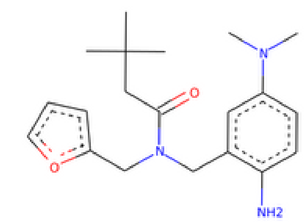
0.6

CN(C)c1ccc(N)cc1CN(Cc1ccco1)C(=O)CC(C)(C)C



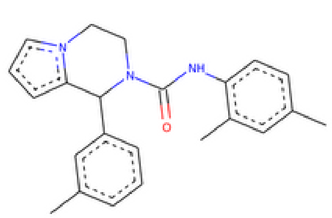
21.0

CN(C)c1ccc(CN(Cc2ccco2)C(=O)CC(C)(C)C)cc1N

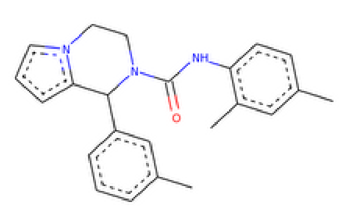


40.1

CN(C)c1ccc(N)c(CN(Cc2ccco2)C(=O)CC(C)(C)C)c1

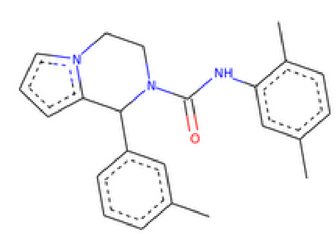


Cc1cccc(C2c3cccn3CCN2C(=O)Nc2ccc(C)cc2C)c1



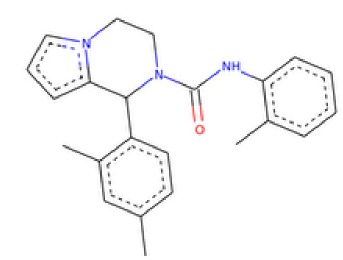
0.5

Cc1cccc(C2c3cccn3CCN2C(=O)Nc2ccc(C)cc2C)c1



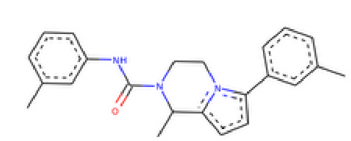
28.4

Cc1cccc(C2c3cccn3CCN2C(=O)Nc2cc(C)ccc2C)c1



31.9

Cc1ccc(C2c3cccn3CCN2C(=O)Nc2ccccc2C)c(C)c1



34.7

Cc1cccc(NC(=O)N2CCn3c(-c4cccc(C)c4)ccc3C2C)c1

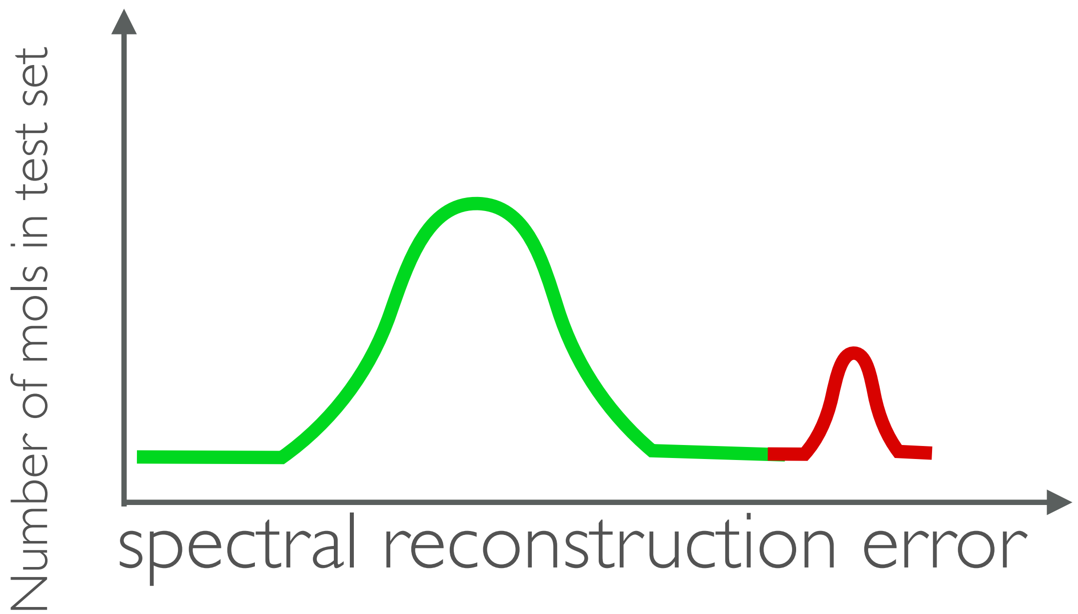
Correct structure highlighted in green

TRADEOFF CURVE

For each molecule in the test set, we use the candidate structure with lowest spectral error as our "guess"

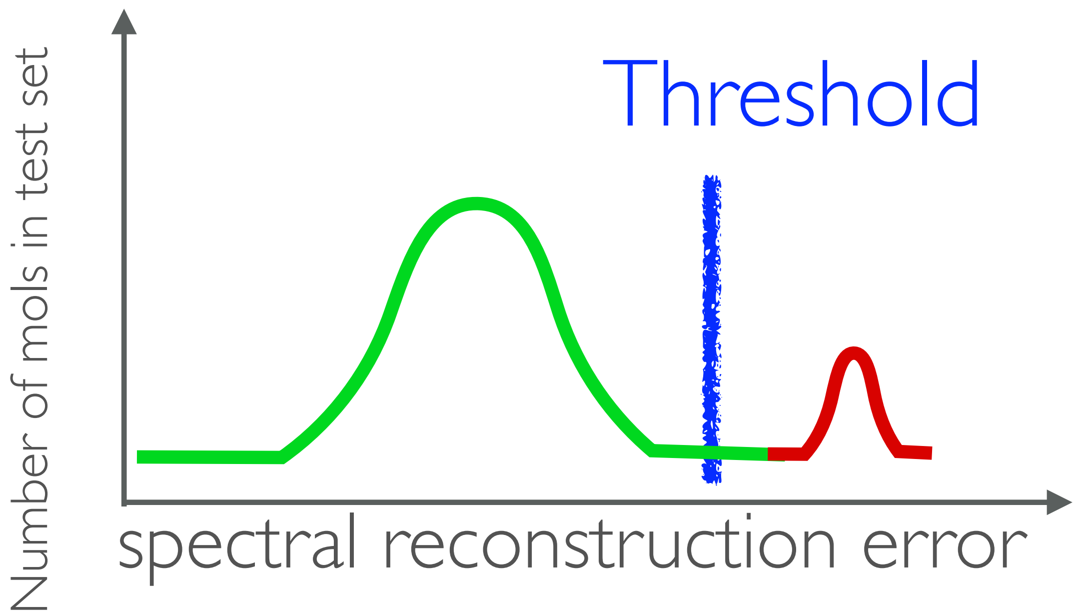
TRADEOFF CURVE

For each molecule in the test set, we use the candidate structure with lowest spectral error as our "guess"



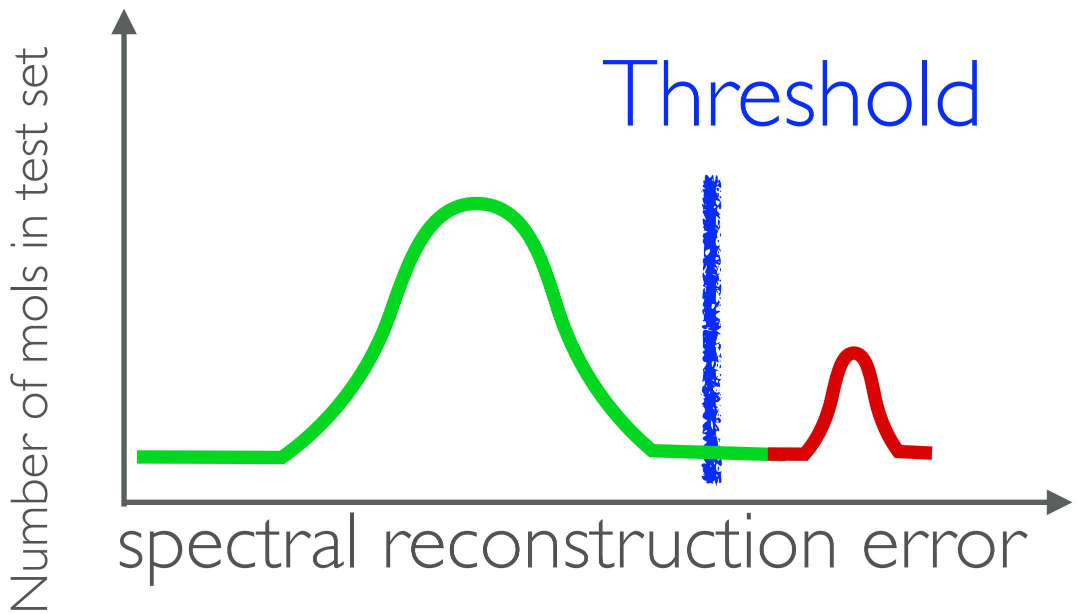
TRADEOFF CURVE

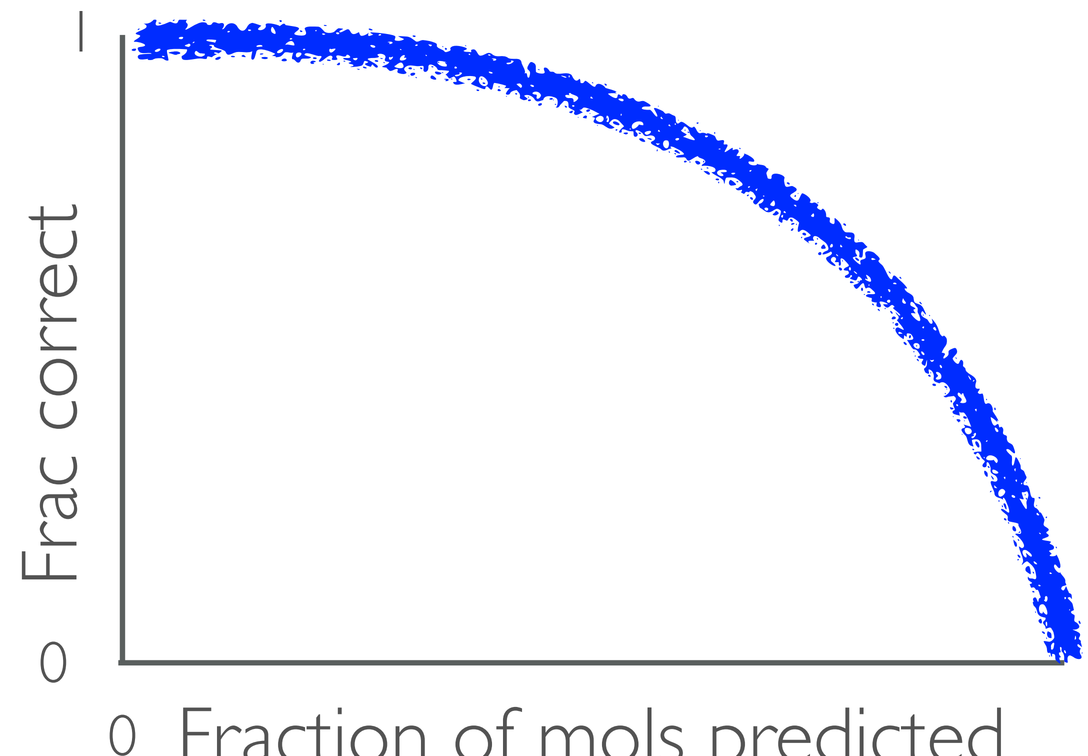
For each molecule in the test set, we use the candidate structure with lowest spectral error as our "guess"



TRADEOFF CURVE

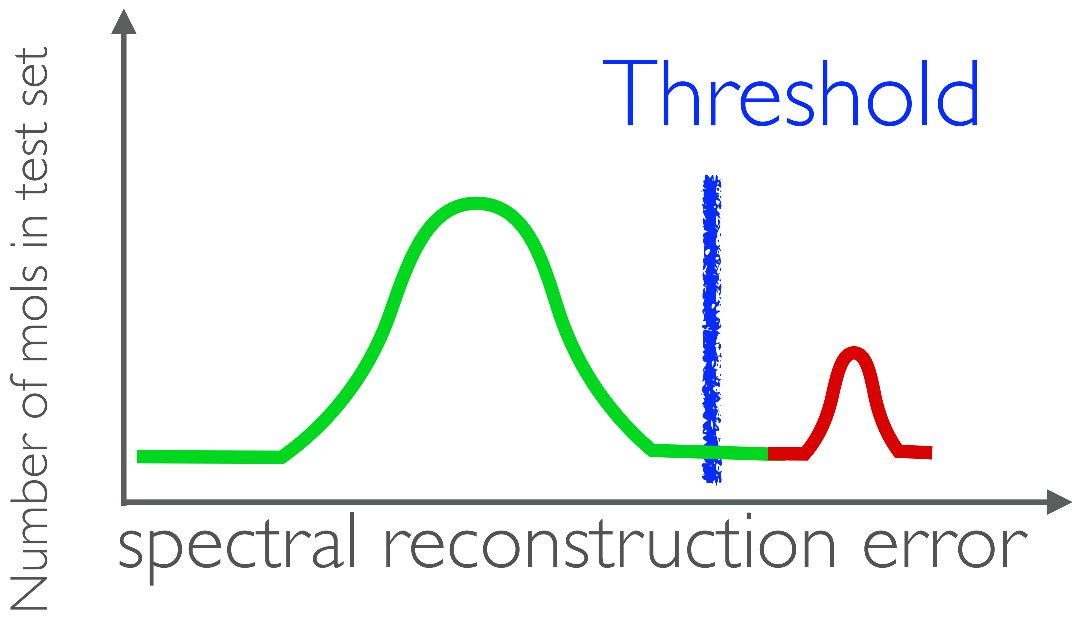
For each molecule in the test set, we use the candidate structure with lowest spectral error as our "guess"

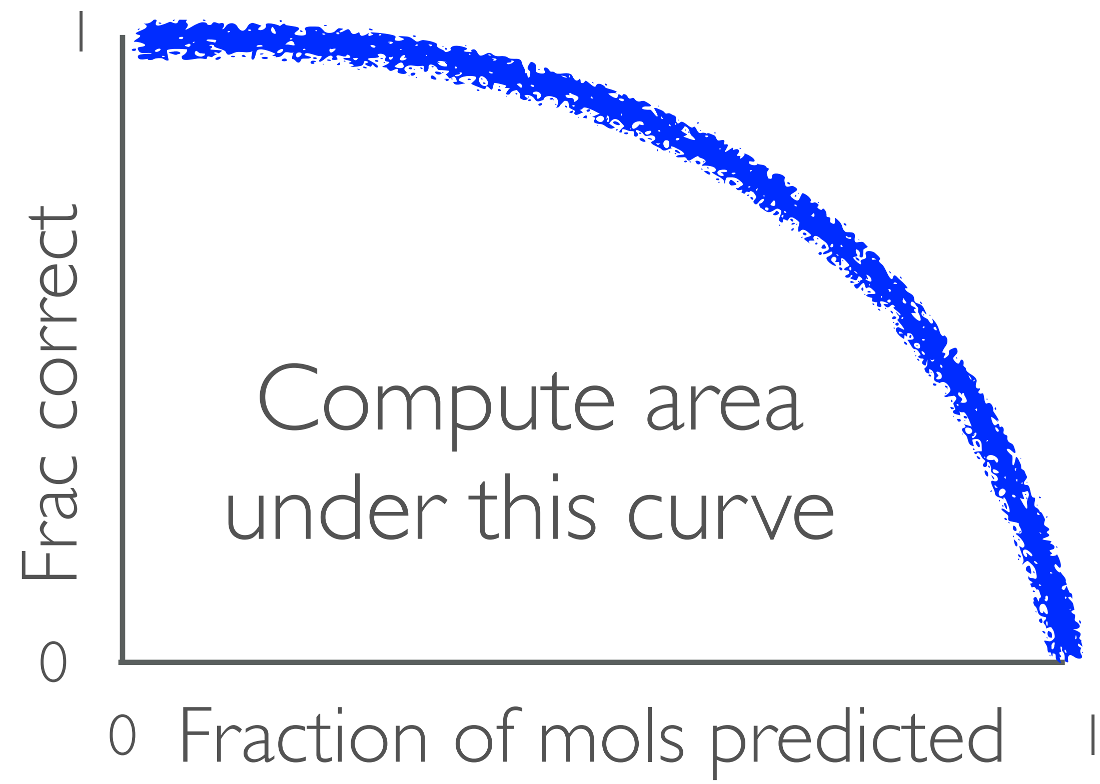




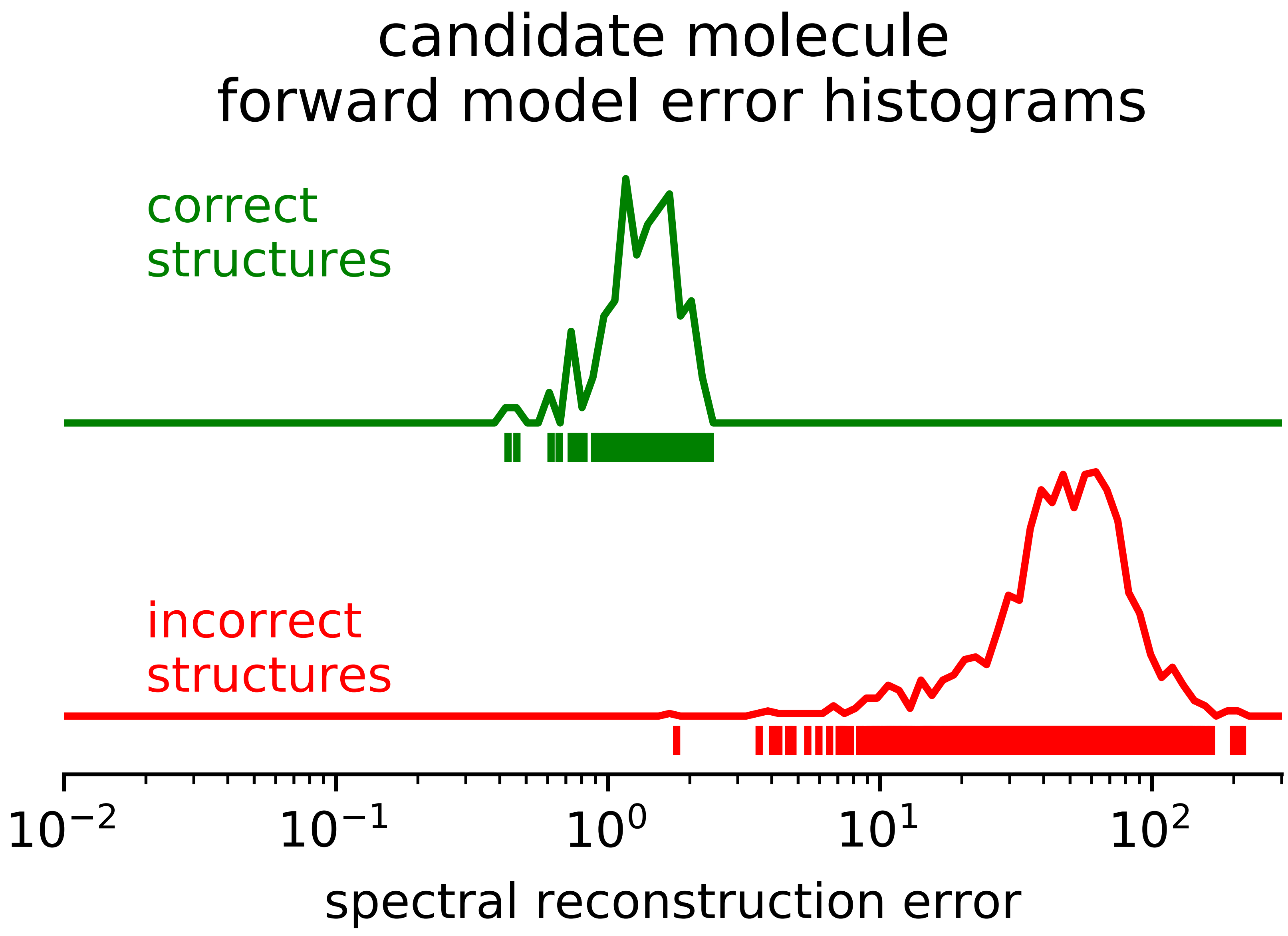
TRADEOFF CURVE

For each molecule in the test set, we use the candidate structure with lowest spectral error as our "guess"





TRADEOFF CURVE



EXPERIMENTAL DATA

Now we're going to use real observed spectra (NOT predicted from our forward model)

Forward Model

Train

Test

Inverse Model

Train

Test

EXPERIMENTAL DATA

Now we're going to use real observed spectra (NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

spectrum agreement percentile

EXPERIMENTAL DATA

Now we're going to use real observed spectra (NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

			spectrum agreement percentile			

EXPERIMENTAL DATA

Now we're going to use real observed spectra
(NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

			spectrum agreement percentile			
Foward	Inverse	AUC	10%	20%	50%	Top-1

EXPERIMENTAL DATA

Now we're going to use real observed spectra (NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

			spectrum agreement percentile			
Foward	Inverse	AUC	10%	20%	50%	Top-1
Train	Train	0.95	100%	100%	99.2%	79.5%

EXPERIMENTAL DATA

Now we're going to use real observed spectra
(NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

			spectrum agreement percentile			
Foward	Inverse	AUC	10%	20%	50%	Top-1
Train	Train	0.95	100%	100%	99.2%	79.5%
	Test	0.88	100%	99%	96.9%	71.3%

EXPERIMENTAL DATA

Now we're going to use real observed spectra
(NOT predicted from our forward model)

- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

			spectrum agreement percentile			
Foward	Inverse	AUC	10%	20%	50%	Top-1
Train	Train	0.95	100%	100%	99.2%	79.5%
	Test	0.88	100%	99%	96.9%	71.3%
Test	Train	0.91	100%	100%	97.6%	60.3%

EXPERIMENTAL DATA

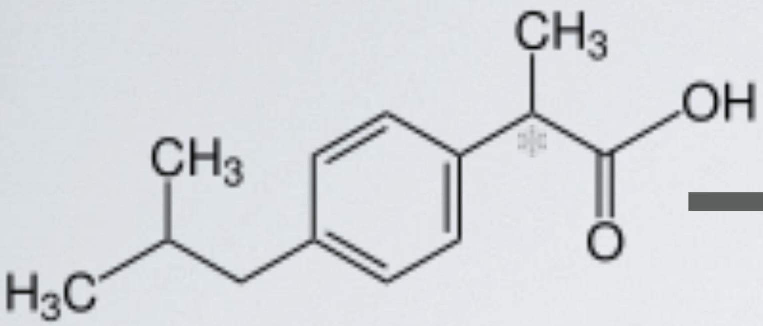
Now we're going to use real observed spectra
(NOT predicted from our forward model)

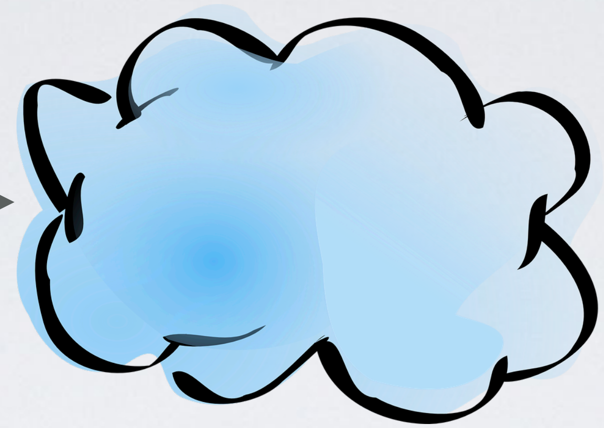
- Train on 1.3M molecules in pubchem
- HCON, up to 32 heavy atoms
- 13C chemical shifts

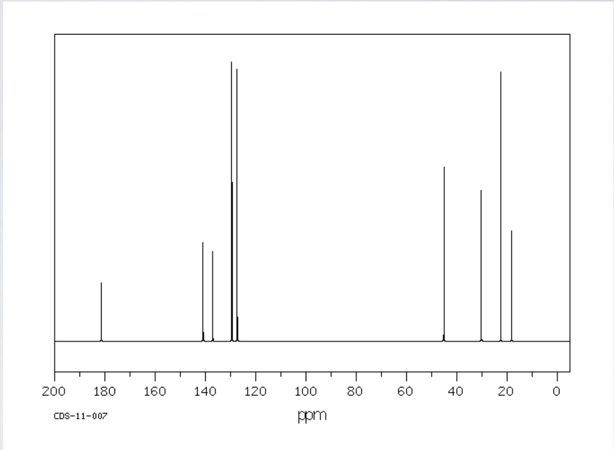
Foward	Inverse	AUC	spectrum agreement percentile: 10%	20%	50%	Top-1
Train	Train	0.95	100%	100%	99.2%	79.5%
	Test	0.88	100%	99%	96.9%	71.3%
Test	Train	0.91	100%	100%	97.6%	60.3%
	Test	**0.83**	**100%**	**97.1%**	**90.2%**	**55.9%**

This is the most realistic scenario — neither the forward or the inverse model had ever seen this structure

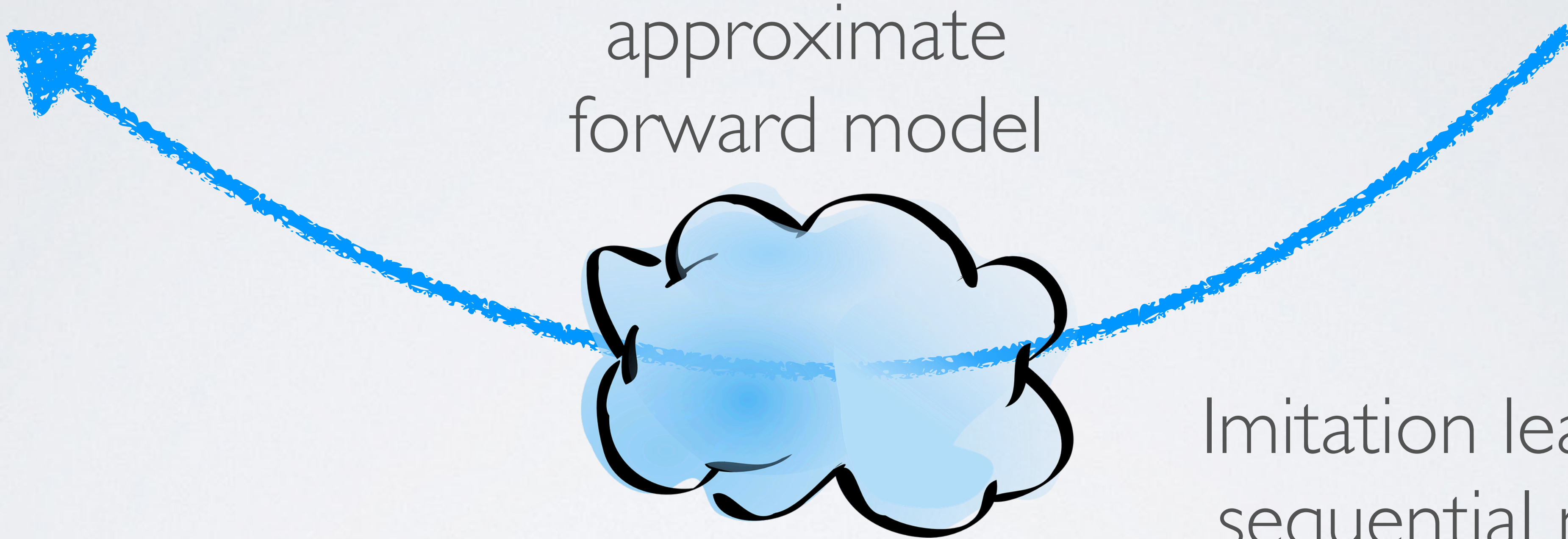
NMR INVERSE PROBLEM





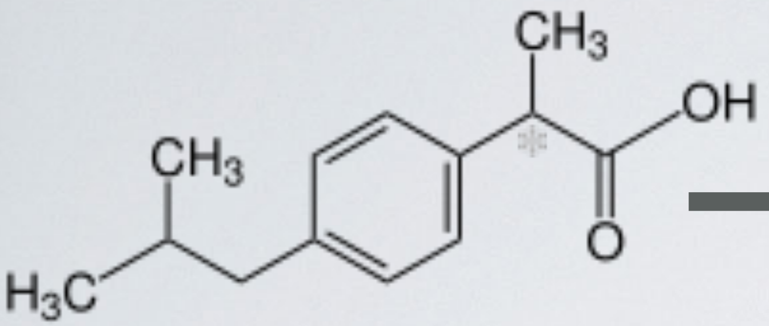


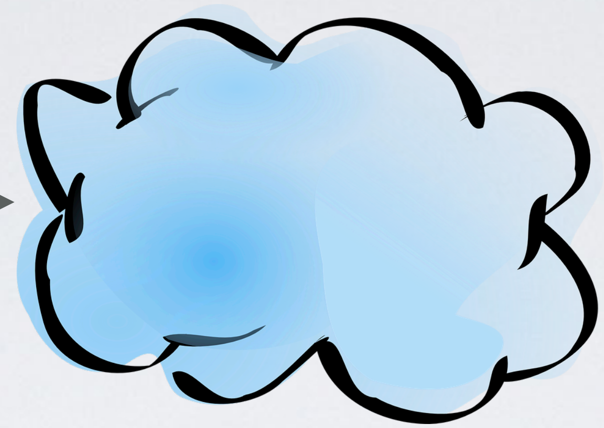
approximate
forward model

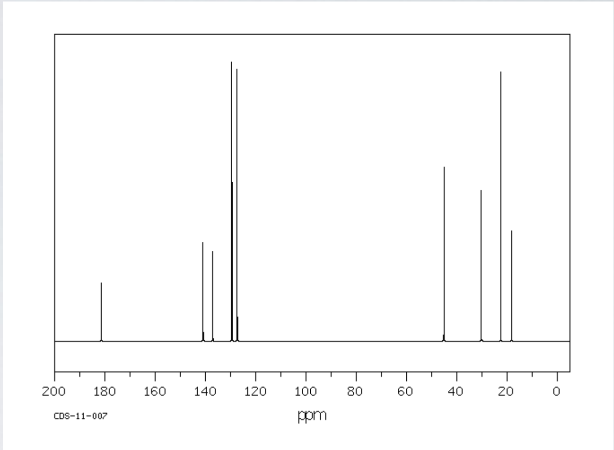


Imitation learning for
sequential molecule
construction

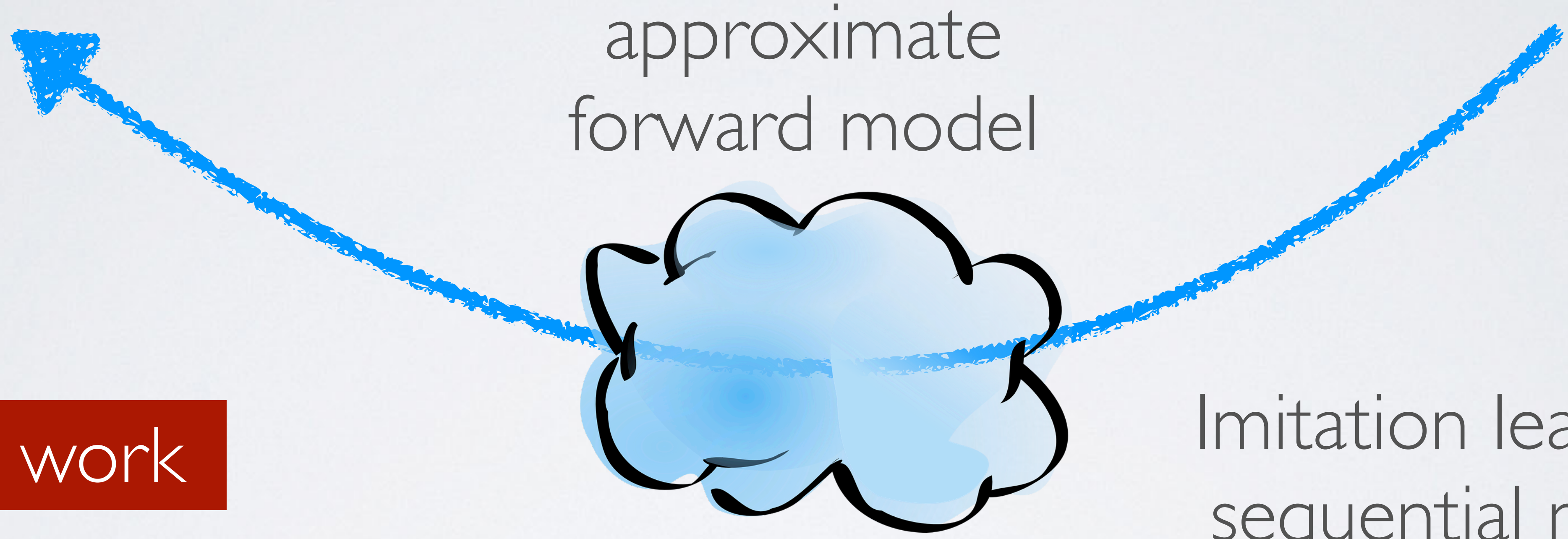
NMR INVERSE PROBLEM







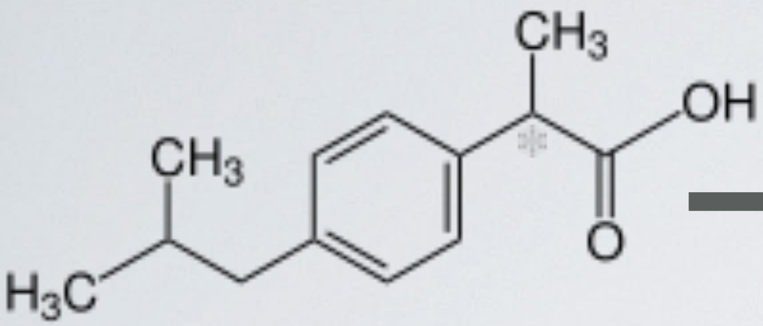
approximate
forward model

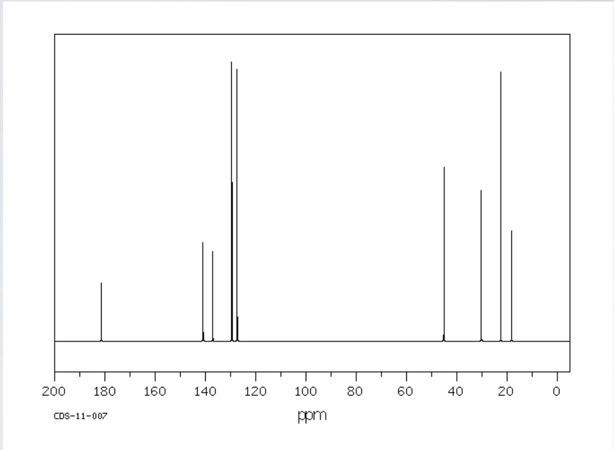


Ongoing work

Imitation learning for
sequential molecule
construction

NMR INVERSE PROBLEM





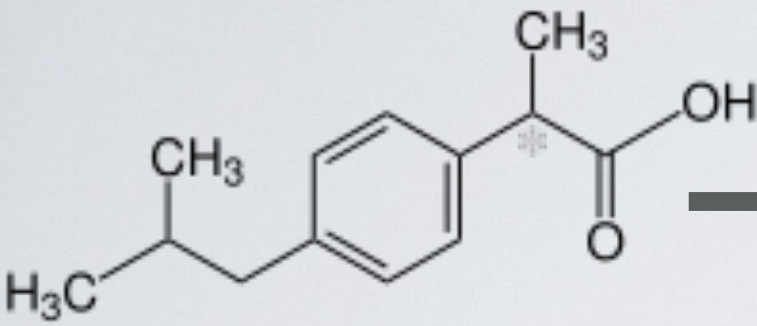
approximate forward model

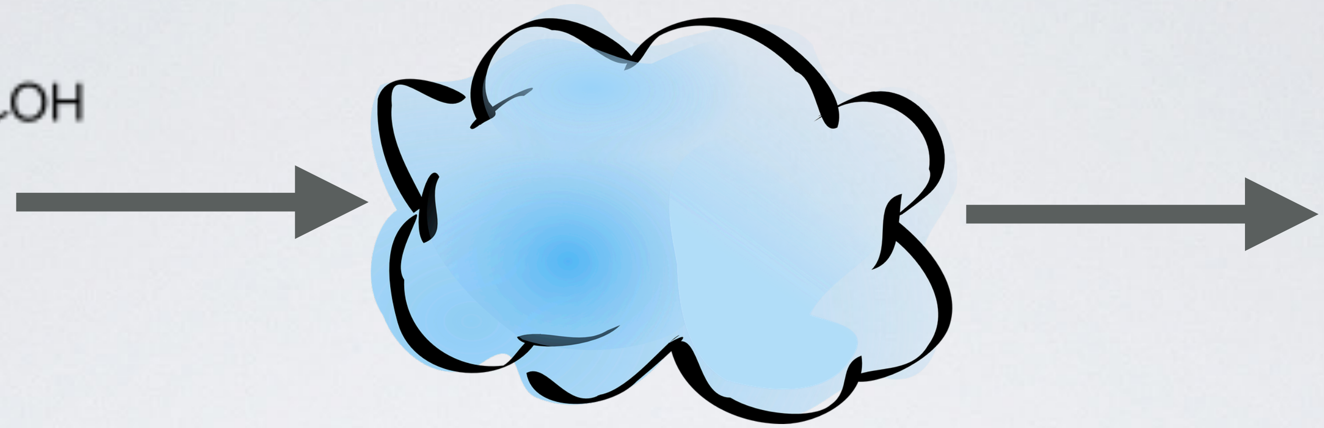
Ongoing work

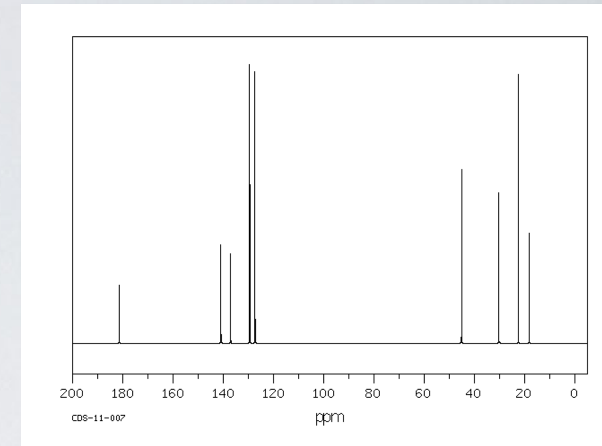
- Increase reconstruction accuracy by correcting for distributional shift

Imitation learning for sequential molecule construction

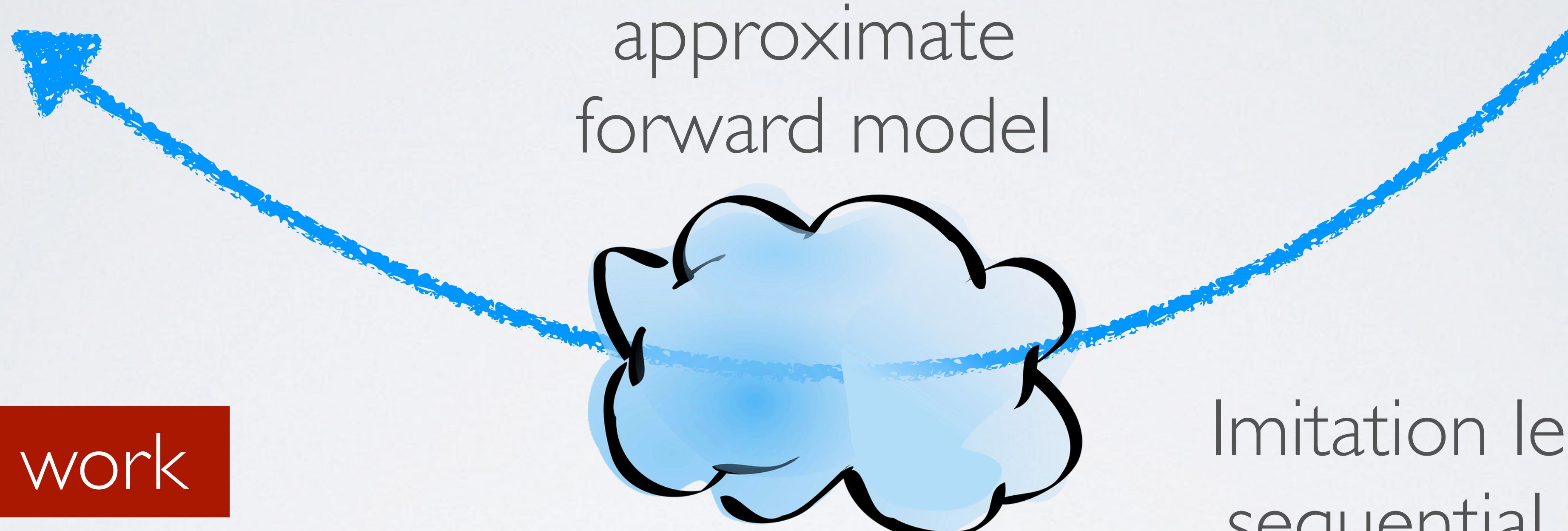
NMR INVERSE PROBLEM







approximate
forward model

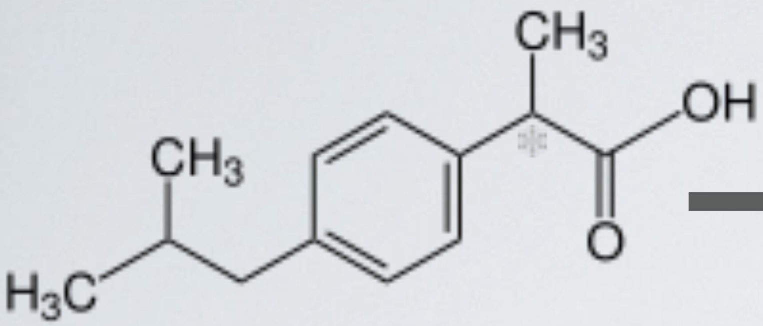


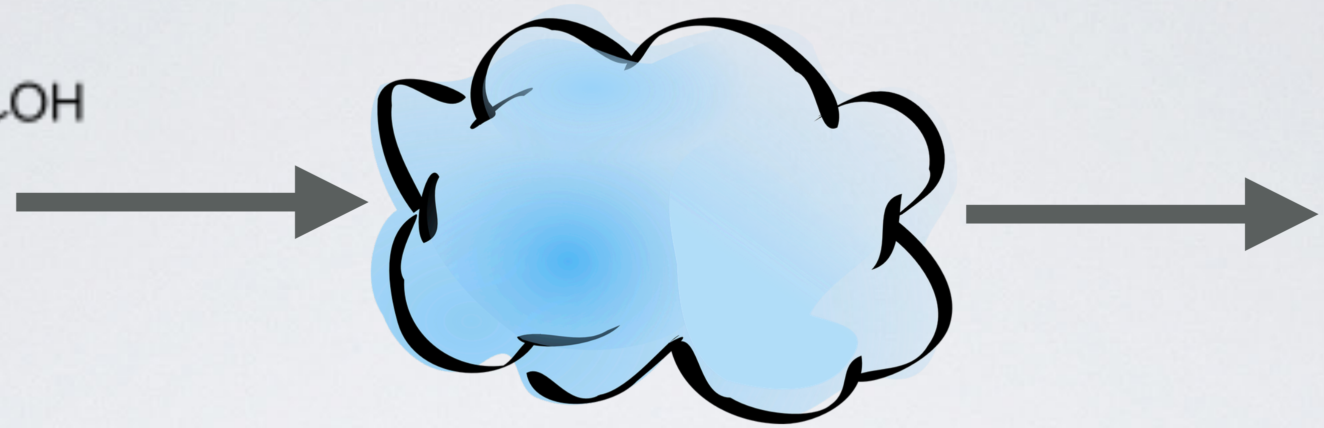
Imitation learning for
sequential molecule
construction

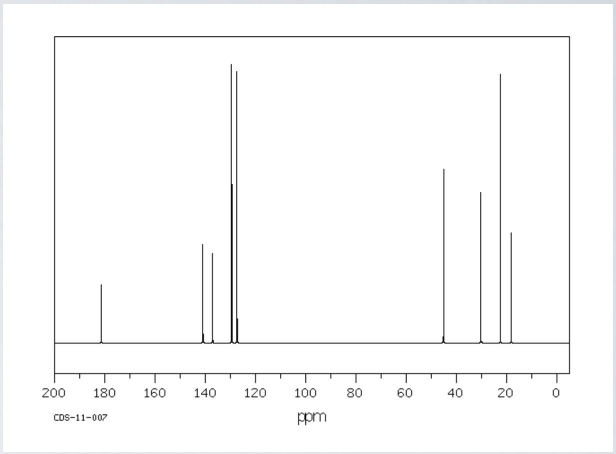
Ongoing work

- Increase reconstruction accuracy by correcting for distributional shift
- Handle stereoisomerism and geometry

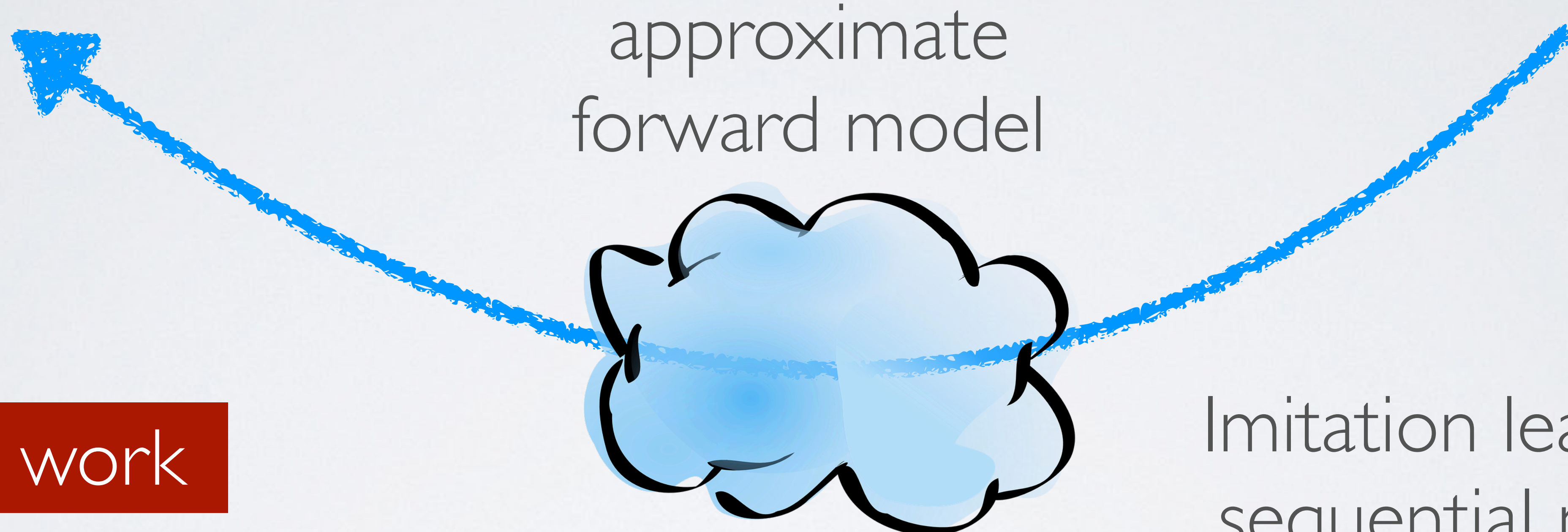
NMR INVERSE PROBLEM







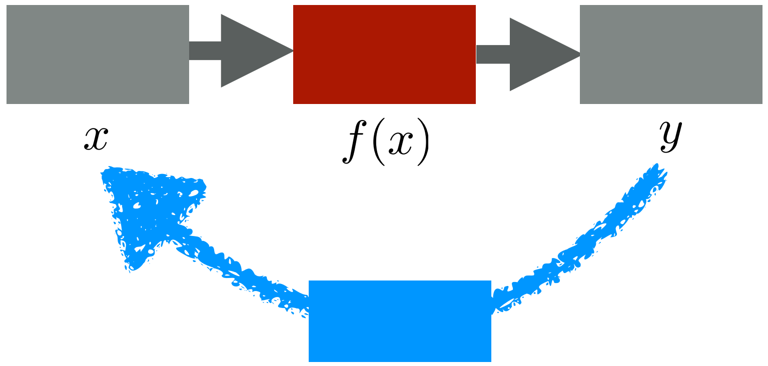
approximate
forward model

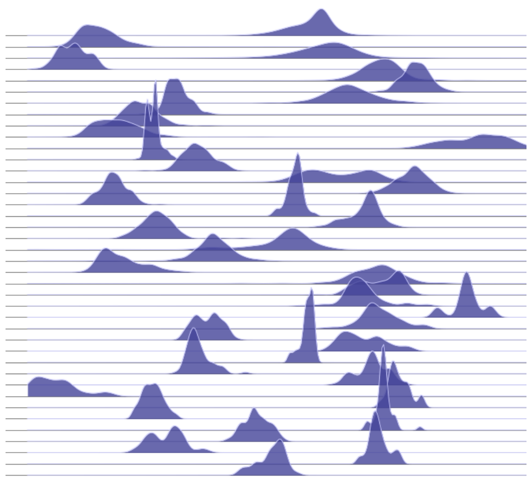


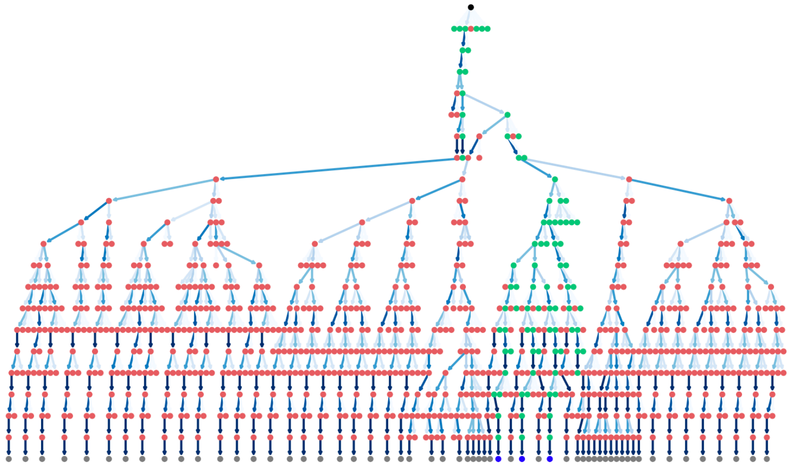
Imitation learning for
sequential molecule
construction

Ongoing work

- Increase reconstruction accuracy by correcting for distributional shift
- Handle stereoisomerism and geometry
- Accelerate reconstruction speed







What are
inverse problems?

Spectroscopy:
The forward problem

Spectroscopy:
The inverse problem

molecular inverse problems

Spectroscopy and RDKit

Eric Jonas

Department of Computer Science
University of Chicago

@stochastician
ericj@uchicago.edu

THANKS